Universität Zürich UZH

Masterarbeit
zur Erlangung des akademischen Grades
Master of Arts
der Philosophischen Fakultät der Universität Zürich

(Key) Phrase Extraction with Masked Language Models and Information Theory

Verfasser: Niclas Bodenmann
Matrikel-Nr: 17-700-436

Referent: Prof. Dr. Rico Sennrich

Institut für Computerlinguistik

Abgabedatum: 01.01.2024

Abstract

This thesis conceptually replicates an n-gram language modeling approach to keyphrase extraction from previous work. It translates the original causal paradigm to a masked language model and qualitatively evaluates the results. The key idea is to compare a token's probabilities given differently masked contexts. This thesis shows how and why the attempted replication is unsuccesful, but finds that the 'phraseness' calculation is highly interesting in terms of idiomaticity. Further experiments geared towards phrasematics confirm not only how the presented probing strategy is able to identify non-compositional idioms, but how it could be used as building block for identifying idiomatic syntactical structures.

Zusammenfassung

Die vorliegende Arbeit präsentiert die konzeptionelle Replikation eines Keyphrase Extraction Ansatzes, welches im Original auf n-Gram-Sprachmodellen basiert. Sie übersetzt die primären Ideen in die Logik eines maskierten Sprachmodells und evaluiert die Resultate qualitativ. Schlüsselidee ist es, die Wahrscheinlichkeiten eines Tokens in unterschiedlich maskierten Kontexten zu vergleichen. Obwohl die Replikation als solche scheitert und die Gründe dafür erläutert werden, stellt sich heraus, dass der präsentierte Ansatz dazu in der Lage ist, idiomatische Ausdrücke zu identifizieren. In zusätzlichen Experimenten wird diese Einsicht bestätigt, indem mitunter gezeigt wird, dass die Methode selbst syntaktische Idiome erkennt.

Acknowledgement

I want to thank Prof. Dr. Rico Sennrich for his supervision and helpful advice, as well as Dr. Jannis Vamvas for helping me understand the technical intricacies of SwissBERT. Special thanks go to Andreas Säuberli, who was an invaluable sparring partner during all phases of work.

Contents

List of Figures

List of Tables

List of Acronyms

IR Information Retrieval. 1, 4, 21

KE Keyphrase Extraction. iv, 1–5, 7, 12, 15, 18–21, 28, 33, 35, 39, 42–44, 53, 68, 69

ML Machine learning. 8, 12

MLM Masked Language Model. iv, 1, 2, 24, 40, 41, 49, 53–55, 68, 70

NLP Natural Language Processing. iii, 2, 7, 12, 13, 15, 19, 20, 68

SMD Schweizer Mediendatenbank. 36, 38

1 Introduction

1.1 Motivation

Keyphrase Extraction (KE) is the process of automatically identifying phrases that are somehow indicative for a certain text or body of texts. As such, KE is of critical importance in helping humans and search engines to navigate the evergrowing mountains of texts. This is not only true for Information Retrieval (IR), but also for researchers such as corpus linguists who directly study these texts and rely on computational methods to interact with them.

This thesis examines an older approach to KE by Tomokiyo and Hurst [2003] and tries to translate it to modern language modeling technology, a Masked Language Model (MLM) to be specific. To do so, the foundations of Information Theory are presented, as well as research that went into keyphrase and keyword extraction and the evaluation of different approaches. Finally, the results of the adaptation are presented and discussed, and further experiments based on these results are conducted.

This is an exploratory thesis. As such, it is not directly interested in producing the best possible results, but to understand how and why the presented approaches work as they do. In doing so, this thesis might show indications for further research that relies on probing an MLM. The code used is available on GitLab.[1]

Finally, I am approaching this thesis with the interests of data-driven, inductive corpus linguistics. Opposed to many KE approaches, I am not particularly interested in thematic or topical key-(noun-)phrases for single documents, but in patternized language use in whole document collections or corpora. These ‘keypatterns’ may be constituted by noun phrases, or by any other sequence of tokens.

[1] https://gitlab.uzh.ch/niclaslinus.bodenmann/master-thesis-code-repository

1.2 Research Questions

The research questions that shall be answered in this thesis, are:

1. How can the Keyphrase Extraction approach by Tomokiyo and Hurst [2003] be translated to masked language models?
2. How does the adapted approach perform for Keyphrase Extraction, and what explains this performance?
3. How can the methodology developed for Keyphrase Extraction be used outside of this context?

1.3 Thesis Structure

Chapter 2 on related work presents unsupervised KE approaches that are note based on the statistical or language modeling methodologies that this thesis revolves around.

In Chapter 3 on theory, the first section introduces fundamental formulae of Information Theory and tries to offer interpretations thereof. The second section gives an overview of different conceptualizations of what constitutes keyphrases. Because the approach I am adapting strives to deliver keyphrases for a whole corpus, as opposed to a single document, both interpretations from corpus linguistics as well as applied Natural Language Processing (NLP) are discussed and contrasted. This section will also serve to present the terminological and conceptual tools that will be used to qualitatively discuss the results of my experiments. The next section on language models will briefly introduce language models as a general concept and compare the two different paradigms of masked language modeling and causal language modeling. It furthermore offers a short history of technical innovations that lead to up to SwissBERT [Vamvas et al., 2023], the model used for my adaptation. In Section 3.4, I will present the original paper "A language model approach to keyphrase extraction" by Tomokiyo and Hurst [2003] The last theory section (Section 3.5) discusses the different ways in which KE can be and has been evaluated. It also defines in which ways I will be drawing from these traditions. Chapter 4 will present the data that has been used to pretrain *SwissBERT*, as well as the data that I am using as a foreground or target corpus.

There are two main experiments that have been undertaken: The first experiment (Section 5.1) is constituted by the conceptual replication of Tomokiyo and Hurst [2003] with an MLM. I define the methodology, explain where and why I needed to make resourceful concessions, present the results and interpret what they mean for the success or failure of the replication. The second experiment (Section 5.2) is motivated by the conclusions of the first, and refocuses on phraseness and idiomaticity. Once again, I present the methodology, potential biases, and discuss the results. For both experiments, the results are curated in the sense that I undertook many more experiments than directly reported, with slightly different parameter configurations and formulaic variations. Because the results needed to be manually assessed and I could not rely on numeric metrics, 'pruning' as necessary.

Lastly, I will conclude the thesis by explicitly answering the research questions and widening the scope of the findings to a broader horizon than KE. Further potential research is pointed out, both within the confines of my approach and outside of it.

2 Related Work

KE is of interest to both industry and academia. In applied settings, it plays a role as a component of both IR and text summarization. Corpus Linguistics on the other hand is interested in KE as a method to arrive at data-driven corpus characterizations.

Due to the breadth of Keyphrase Extraction (KE) research and the focus of this thesis, I will center this section around unsupervised KE. Most often, unsupervised KE is conceptualized and implemented as a two/three-step-process [Hasan and Ng, 2010], consisting of *Candidate selection*, *Candidate ranking* and, depending on the candidates, *Phrase formation.*

Papagiannopoulou and Tsoumakas [2020] name four different methodological approaches to unsupervised KE.

1. Statistics-based;
2. Graph-based;
3. Embeddings-based;
4. Language model-based.

I will present here the basic ideas behind graph- and embeddings-based models. Statistics- and language model-based approaches will be discussed in Chapter 3.

Graph-based approaches build graphs based on candidate phrases as nodes and apply an adjusted *PageRank* [Page et al., 1999] algorithm to identify keyphrases. *PageRank* was introduced as a ranking algorithm for websites in the context of information retrieval. Each page or document is assigned a score based on how many other pages link to it. The score is calculated recursively, links from pages with high scores are considered more important. *TextRank* by Mihalcea and Tarau [2004] was the first KE approach to employ graphs and an adaptation of *PageRank.* After preprocessing the text and selecting candidate phrases based on a syntactic filter, they treat the phrases as nodes, and draw edges between the nodes based on cooccurence. Compared to the original *PageRank*, their approach therefore worked with an undirected graph. Later research extended this approach by edge weights based on the frequency of cooccurence [Wan and Xiao, 2008].

Other approaches based on graphs integrate information from other documents than the target document, but which are connected to it, e.g via citations [Wan and Xiao, 2008; Gollapalli and Caragea, 2014].

In 2009, Liu et al. [2009] were the first to extract keyphrases based on clusters, although not yet graph-based. A year later, Liu et al. [2010] extended the idea to a graph-based algorithm that considered the topics of a document, obtained via LDA. Bougouin et al. [2013] use agglomerative clustering to assign candidate phrases to different topics. Later topic-based KE refined these methods by introducing multipartite graphs [Boudin, 2018] or by proposing more resource-efficient algorithms [Sterckx et al., 2015]. The central idea behind introducing topics is that keyphrases should not be regarded on their own, but in context with all other keyphrases extracted for a document. The keyphrases should cover the thematic spread of a document, and not only the main idea of it. If for example three keyphrases should be extracted for a document with three central topics, there should be one keyphrase for each topic. Lastly, several KE draw upon graphs that are constructed between candidate phrases and knowledge graphs. Shi et al. [2017] employ such an approach by integrating DBpedia, or Yu and Ng [2018], by connecting candidates to Wikipedia.

The usage of embeddings is both used in conjunction with graphs and subsequent *PageRank*-inspired algorithms [Wang et al., 2015; Mahata et al., 2018], or as a standalone approach. Bennani-Smires et al. [2018] embed the candidate phrases and the document and choose phrases with the highest cosine similarity to the document. Papagiannopoulou and Tsoumakas [2018] train embeddings on single target documents, and then rank keyphrases based on their similarity to the vector representation of title and abstract of the document.

Unsupervised KE techniques that rely on the transformer architecture and pretrained language models have started to gain traction. *AttentionRank* [Ding and Luo, 2021] extract the accumulated self-attention of *BERT* [Devlin et al., 2019], i.e. the attention a candidate phrase receives, as well as the cross-attention between the candidate phrase and the sentence it appears in to determine the importance of a candidate phrase. *MDERank* [Zhang et al., 2023] compares the *BERT* document embedding with the same embedding once candidate phrases are masked, and determine a candidate's importance based on the resulting difference. *MDERank* relies on finetuning *BERT* to obtain *KPEBERT*, with a newly introduced finetuning objective that incentivizes the model to assign differing embeddings when keyphrases are masked. For training, they obtain keyphrases using *YAKE!* [Campos et al., 2018], a lean statistical KE method, which will be discussed in Section 3.2.2. Therefore, *MDERank* is not strictly unsupervised. Similarly, *UCPhrase* [Gu et al., 2021] use 'silver' labels and *BERT* to generate attention maps, which are then used to train a simple sequence labeler. They obtain the silver labels via a heuristic that looks for the longest, repeated token sequences in a document. *PromptRank* [Kong et al., 2023] measures the probability of a candidate based on a prompt. A target document is encoded, and the probability of the candidate phrases is measured for when the decoder is prompted with a prefix such as "The central topic of this document is:". Lastly, the probability of the decoder is combined with a heuristic based on the candidate's position in the document.

3 Theory

This chapter presents and discusses the theoretic foundation of KE as pursued in this thesis. It gives a brief overview of Information Theory in the context of computational linguistics and presents the language model architectures primarily used in this thesis. Furthermore, I give an introduction into keyphrases from both the perspective of applied NLP, corpus linguistics and linguistics in general.

3.1 Information Theory

Information Theory as its own field was developed and codified by Claude Shannon, an engineer at Bell Labs, in the year 1948 with the seminal paper "A Mathematical Theory of Communication" [Shannon, 1948]. Shannon unified and generalized concepts from statistical mechanics (e.g. the 'entropy' of Boltzmann [1872]) and previous works from colleagues at Bell Labs and built the foundation for decades of research across many disciplines.

The Information Theory proposed by Shannon is fundamentally structured by the concept of *communication*. The underlying Shannon-Weaver model of communication posits an information source and a destination, connected by a channel through which a signal, an encoded message, is transmitted. The channel can be subject to noise that is corrupting the signal. The source uses a transmitter, dubbed encoder by later research, to transform the message into a signal that is, ideally, robust to the noise. On the other hand, the destination is using a receiver or decoder to „reconstruct[...] the message from the signal." [Shannon, 1948].

In his introduction, Shannon makes two conceptual points, that I want to highlight. Firstly, he explains how *information* is not to be confused with the semantic contents of a message: "These semantic aspects of communication are irrelevant to the engineering problem" [Shannon, 1948, p.1]. The communication model will prove to be influential even for sciences concerned with such semantic contents, but Shannon designed it with engineering in mind. When talking about keyphrases, this distinction must be kept in mind.

Secondly, Shannon cautions against interpreting the model's components literally. Although inspired and conceptualized by „physical counterparts“ [Shannon, 1948], the components of the model are to be thought of as „mathematical entities“. It is this second argument, or rather, how it guided Shannon's formulations, that allowed for the wide applicability of information theory. Singular measures, formulae and concepts of Shannon's work can be interpreted and used without needing to actualize or account for the whole framework.

For the further discussion of units and formulae, I will refer to the introductory work "Information Theory, Inference, and Learning Algorithms" by MacKay [2003]. Widely cited, MacKay puts Information Theory in a modern context with regards to Machine learning (ML) and neural networks. I will adhere to his notation standards.

Before delving into the formulae, I want to stress the abstract mathematical nature of Information Theory. During my research and writing, I did not find a single conceptual field or interpretation that can account intuitively for all the formulae discussed below.[1] Therefore, I am going to use two different interpretations and want to make them explicit. The first interpretation is the one of the *uncertainty* of the receiver. This interpretation views Information Theory from the perspective of the receiver, according to the model outlined above.[2] The receiver observes a distribution and does not know what outcome will come to be. Depending on the distribution (which is known by the receiver), there is more or less uncertainty. Once the outcome is observed, the uncertainty is reduced. The second interpretation comes from the domain of *data compression*, or coding. Here, on the sender side, we attempt to find an optimal (shortest) code to inform the receiver about the outcome of a random variable. For simplicity, we assume a binary code and a noiseless channel.

Shannon assumes that a concrete message is „*selected from a set* [emphasis in the original] of possible messages“ [Shannon, 1948]. If this set is finite, as is the case with alphabetic symbols or binary events, a message x can be regarded as the outcome of the random variable X with discrete possible values.

Thus, the *Shannon information content* of a message or outcome x is defined as in Equation (3.1) [MacKay, 2003].

$$h(x) = \log_2 \frac{1}{P(x)} \tag{3.1}$$

[1] compare this to the different fields of application listed by Cover and Thomas [2005].

[2] In fact, the Information Theory outlined by Shannon productively conflates the view of the sender and the receiver with regards to the concept of *uncertainty*. For a precise discussion, please consider Cole [1993].

The relationship between the information content and the probability of the message is inverse: The less likely the outcome, the higher the information content.

When regarding this formula in the context of *coding*, the shannon information defines the lower bound for the length of the code when encoding this outcome in binary (when using an unambiguous code for all possible events).[3] Consider an equiprobable event space with four outcomes, each with the probability of 0.25. The information content of each outcome is $\log_2 \frac{1}{0.25} = 2$, and the optimal (in terms of brevity) binary codes for these events are 00, 01, 10 and 11.

Directly related to the information content is the *entropy*, which describes the average information content of an event x from random variable X (Equation (3.2)). As noted in the previous example, the information also describes the lower bound for the shortest possible signal length for a single event, thus the entropy describes the lower bound for the optimal *average* signal length.

$$H(P) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log_2 \frac{1}{P(x)} \tag{3.2}$$

$\mathcal{A}_X$ denotes the alphabet, or possible outcomes of X. Furthermore, it is conventionalized for $P(x) = 0$ that $0 \times \log_2 \frac{1}{0} \equiv 0$, since the whole term trends towards 0 with lower and lower probabilities. The first factor $P(x)$ can be considered as how probable the outcome is, according to distribution P. The second factor $\log_2 \frac{1}{P(x)}$ is the Shannon information or the optimal message length when considering distribution P

In practice, especially in machine learning, we often like to compare different distributions over the same alphabet or event space. This leads to the notion of cross-entropy in Equation (3.3), wherein the probabilities of a second distribution Q determine the information content (or optimal message length), while distribution P still determines how probable the outcomes are.

$$H(P, Q) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log_2 \frac{1}{Q(x)} \tag{3.3}$$

In coding terms, the cross entropy describes the average amount of bits used to encode P when using a code based on Q. The cross entropy of P with any other distribution Q is always larger or equal to the entropy of P, while it is only equal when $Q = P$. This can be intuitively understood when considering that the most efficient code for P, i.e. the code that on average uses the least amount of bits, should be directly based on P.

[3] For a proof consider MacKay [2003, p. 82ff.] on Shannons' source coding theorem.

In the context of machine learning, the loss of a prediction is often measured with the cross entropy (Equation (3.4a)). Consider a model that generates (or predicts) a distribution Q over all possible labels $\mathcal{A}_X$ for an input. The true distribution P on the other side allocates the total probability mass to the one true label x_t. The deterministic distribution P vastly simplifies the cross entropy to Equation (3.4c).

$$\begin{aligned} H_{ML}(P,Q) &\equiv \sum_{x \in \mathcal{A}_X} P(x) \log_2 \frac{1}{Q(x)} && (3.4a) \\ &\equiv 0 \times \log_2 \frac{1}{Q(x_1)} + \cdots + 1 \times \log_2 \frac{1}{Q(x_t)} + \cdots + 0 \times \log_2 \frac{1}{Q(x_n)} && (3.4b) \\ &\equiv \log_2 \frac{1}{Q(x_t)} && (3.4c) \end{aligned}$$

The excess amount of information or bits that comes to be when using Q to approximate or encode P is called the relative entropy, or more prominently, the *Kullback-Leibler divergence* (c3.5a). When simplified, one arrives at the formulation as in Equation (3.5b), often found in the literature.

$$\begin{aligned} D_{\mathrm{KL}}(P \| Q) &= H(P,Q) - H(P) && (3.5a) \\ &= \sum_{x \in \mathcal{A}_X} P(x) \log \frac{P(x)}{Q(x)} && (3.5b) \end{aligned}$$

D_{KL} is 0 when the two distributions are equal, and grows the larger the two distributions differ. It is very important to note that D_{KL} is not symmetric.

So far, different distributions over the same event space have been regarded, whereas an interpretation in the context of *coding* was useful. For the following formulae however, the concept of *uncertainty* comes in handy. MacKay [2003] uses *uncertainty* as another word for entropy, and it describes how uncertain a receiver is about the outcomes of X. In the case of a bent coin, the entropy is highest when the coin is perfectly fair ($H(X) = 1$), and someone observing the experiment (a receiver) is just left guessing with maximum uncertainty. If the coin however is biased in such a way to only ever land on one side, both the entropy and the uncertainty would be zero. A receiver wouldn't even have to witness the experiment to know the outcome.

When considering two different random variables X and Y, their joint entropy can be described by using the joint probabilities of two events x and y as in Equation (3.6).

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log_2 \frac{1}{P(x,y)} \tag{3.6}$$

The conditional entropy $H(X|Y)$ is given in Equation (3.7):

$$H(X|Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log_2 \frac{1}{P(x|y)} \tag{3.7}$$

MacKay [2003, p. 138] describes the conditional entropy as the "average uncertainty that remains about x when y is known". Building upon this, the *mutual information* is understood to be the average reduction of uncertainty about X when knowing Y (Equation (3.8a)). I deviate from MacKay's notation of the mutal information (which he defines as $I(X;Y)$), due to analogy to formulae later introduced later in this thesis.

$$\begin{aligned} MI(X;Y) &\equiv H(X) - H(X|Y) && (3.8a)\\ &\equiv \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log_2 \frac{P(x \mid y)}{P(x)} && (3.8b)\\ &\equiv \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} && (3.8c) \end{aligned}$$

The mutual information is symmetric, meaning $MI(X;Y) = MI(Y;X)$. This becomes clear when the formula in Equation (3.8b) is transformed with Bayes' rules to Equation (3.8c), where the terms containing y and x occur symmetrically.

In applications, one is often interested not in the whole distribution, but how single events inform each other. This concept of pointwise mutual information goes back to [Fano, 1961, p. 27], not yet named as such, but formulated as in Equation (3.9a). In later literature, the symmetric formulation in Equation (3.9b) is most often found.

$$\begin{aligned} pMI(x;y) &= log_2 \frac{p(x \mid y)}{p(x)} && (3.9a)\\ &= \log_2 \frac{p(x,y)}{p(x)p(y)} && (3.9b) \end{aligned}$$

3.2 Keyness, Phrases, Keyphrases

The literature review articles I have considered to find an entrypoint to keyphrase extraction use NLP or ML methods to classify and structure the wide range of approaches to keyphrase extraction. I will not be following this tradition. Instead, I will try to identify the underlying modeling ideas behind different approaches and machine learning features. The following explanations draw from two research traditions, KE as practiced in NLP, and keyword or keyness analysis in corpus linguistics.

First, I will consider how linguistics and especially corpus linguistics have discussed notions of keywords and phrases. This will provide a conceptual and terminological framework to discuss the methods, approaches and results of the NLP community.

3.2.1 Keyness and Phrases in (Corpus) Linguistic Research

3.2.1.1 Unlocking Texts with Key Items

The term 'keyness' shall broadly cover the quality of an item, word or phrase, that makes it a 'key' item. Depending on the approach or tradition, different aspects can make an item 'key', and I will try to give a comprehensive overview below. While many concepts and methods overlap, the NLP community is usually concerned with the keyness of terms in the context of a document, while corpus linguistics regards keyness usually for whole corpora or document collections.

In corpus linguistic papers and monographies, keywords are usually introduced as either being exemplary for the "aboutness" of a corpus, or especially in earlier research, starting with Scott [1997, p. 236], as "word[s] that occur[s] with an unusual frequency in a given text."[4] The first definition is broad enough to house all possible implementations and interpretations, while the second one leaves linguistic or language theoretic considerations aside. Research has gone into fleshing out the conception(s) of keyness, and especially for the purposes of evaluating and discussing the results obtained in the experiments, different aspects will be introduced here.

Stubbs [2010] examined and structured previous (linguistic and humanistic) keyword research into three categories. The first refers to one of the earliest uses of the english term "keyword" in academia by Williams [1985]. For Williams [1985, p. 15], keywords are culturally significant words, "indicative words in certain forms of thought". They are not calculated, but identified by the hermeneutic reading of culture and discourse,

[4] For a discussion of research that has been undertaken in the field before Scott [1997] has used the term "key word", please consider Gabrielatos [2018].

and include terms such as "Bourgeois" and "Rational" [Williams, 1985, p. 45 and p. 252]. Stubbs [2010] shows how German and French scholars and lexicographers came up with this notion themselves during the beginning and middle of the 20th century under the terms "Schlagworte" and "mots-clés" respectively.

The second sense of keywords is directly related to their algorithmic and quantitative definition, defined by Scott [1997] and popularized by its implementation in the *WordSmith Tools* corpus software [Scott, 1998]. Stubbs [2010] criticizes these definitions for how monolithic the corpora are regarded, without a consideration of document boundaries. This will be addressed by later research under the term *dispersion*.

The last sense of keywords Stubbs [2010] discusses comes from corpus linguistic and phraseological research by Hunston and Francis [2000]. Their 'Pattern Grammar' posits that lexico-syntactic patterns with variable concrete word forms are semantically fundamental to language, opposing generativist or structuralist views of language. These patterns, or rather the actualizations within corpora can be identified automatically, and Stubbs [2010] views this as the most ambitious interpretation of keywords, or rather, 'keypatterns'.

Another use of keywords (under this term) I did not find reflected in either corpus linguistic or NLP research is in the context of language learning, where it is used twofold. Introduced by Atkinson [1975], the keyword technique is a mnemonic device to remember words in a foreign language by connecting them with words or phrases sounding similar in a known language. The keyword is the word in the known language, granting access to the memory of the foreign word. More closely related to other notions discussed in this thesis are keywords in the context of reading comprehension. In one approach to vocabulary learning, keywords are defined and explained before reading a text in which they appear [Grabe, 2008]. These keywords are central to understanding a text, while also serving as an opportunity to learn other words related to them in context.

In a recent review article for corpus linguistic keyness analysis, Sönning [2023] provides four understandings of keyness, based on the combination of two dimensions (Table 3.1). Sönning [2023] systematicizes older and established approaches that were mainly concerned with the frequency of terms together with newer methods that emphasize dispersion, i.e. the distribution of a term inside a corpus. It is strongly guided by technical feasibility: The axes are defined by algorithmically useable distinctions, while the aspects themselves are semantically interpreted.

Discernibility refers to the quality of a term to stand out among other terms of the foreground, may it be a single document or a corpus of texts. This is usually translated to a high frequency in the foreground corpus, which is a sensible operationalization. For

	Frequency-oriented	Dispersion-oriented
Target variety in isolation	Discernability of item in the target variety	Generality across texts in the target variety
Comparison to reference variety	Distinctiveness	Comparative generality relative to the reference variety

Table 3.1: Four aspects of keyness according to Sönning [2023]

single documents however, words that appear only once, so called hapax legomena, are themselves standing out, and may even be topic markers [Kew et al., 2023]. With multimodality arriving in linguistic research, typography and layouting may also inform discernibility.

Distinctiveness comes from the original definition by Scott [1997], the comparison of term frequency between two corpora. In its abstract or conceptual form, distinctiveness relates back to findings of Gestalt psychology and its principle of figure-ground perception. It posits that for the perception of a figure, the existence of a background, from which the figure can be distinguished, is fundamental [Peterson and Salvagio, 2010]. In the context of keywords or keyphrases, the distinctiveness informs on how well an item distinguishes the foreground corpus from the background corpus. Depending on the background, different items or features of the foreground become distinctive.

This becomes clear when considering Figure 3.1, where the same gray circle is compared to different backgrounds. On the left side, the color "gray" makes the foreground item distinctive, while on the right side it is the shape "circle".

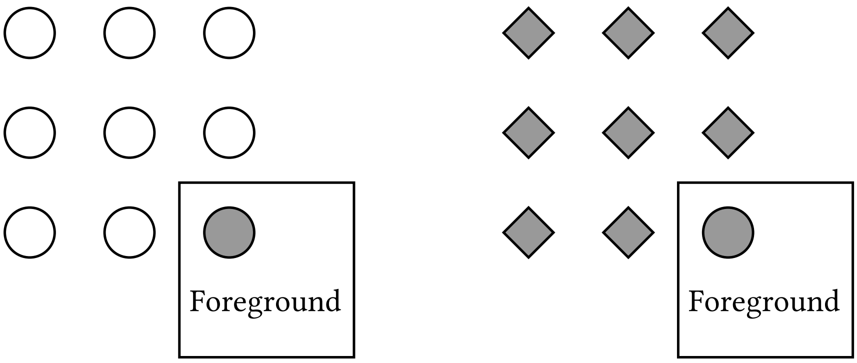


Figure 3.1: Same foreground with different backgrounds

Generality once again only focuses on the foreground or target. It is based on more recent research that identifies the quantitatively assessable *dispersion* as an important measure when discussing keywords. Starting with Baker [2004], who determined the document frequency of terms and set a minimum threshold, measures for dispersion became increasingly refined [Egbert and Biber, 2019; Gries, 2021]. The motivation for the inclusion of dispersion into keyness calculations came from empirical observations. Kilgarriff [1997] noted how the word “whelk” (a carnivorous sea snail) can appear many times in a single document about whelks, but nowhere else in the corpus, making it a bad keyword for the corpus as a whole. Generality thus refers to the quality of a term to be a representative of the whole corpus.

Comparative Generality is analogous to the comparison of a term’s dispersion across two corpora. A keyword of a foreground corpus should be more dispersed in said corpus than in the background corpus. Consider the comparison of a movie discussion corpus and a corpus compiled from NLP papers. In both corpora, the word “Transformer” appears with the same overall frequency. In the movie corpus due to the Transformers movie franchise, and in the NLP corpus because researchers often refer to the foundational architecture in their related work section. However, because only a handful of movies are about Transformers, and the discussions of other movies never mention the term once, whereas many of the NLP papers mention the transformer, “Transformer” would have a higher keyness for the NLP corpus in terms of comparative generality.

3.2.1.2 Methods of Corpus Linguistics towards Keywords

Corpus linguistics has brought forth a suite of methods to automatically identify ‘keywords’ of corpora, outlined in Figure 3.2. They mainly rely on statistical tests for association, which build upon contingency tables. For two binary (present/not present) variables, the observed frequencies are compiled and expected frequencies are derived from that, based on a null hypothesis of independence. Consider Table 3.2 and Table 3.3 for how to calculate and conceptualize the values, taken from Stefanowitsch [2020].

For keywords, A or ¬A is a stand-in for the candidate word, whereas B and ¬B denote the foreground or background corpus respectively. O_{11} is then the count of the candidate in the foreground, O_{21} the count of every other word in the foreground, and so forth. Although originated for unigrams or single words, corpus linguists also use the same method to identify key n-grams (e.g. Bubenhofer [2017]), arriving at something one could also consider KE. However, due to the differing scope and research interest, I would refrain from conflating corpus linguistic keyword identification with KE in the applied setting. Corpus linguistics frequently employ the same methodologies for collocation

	B	¬ B	Total
A	O_{11}	O_{12}	O_{1T}
¬ A	O_{21}	O_{22}	O_{2T}
Total	O_{T1}	O_{T2}	O_{TT}

Table 3.2: Contingency table: Observed values

	B	¬ B	Total
A	$E_{11} = \frac{O_{1T}}{O_{TT}} \times O_{T1}$	$E_{12} = \frac{O_{1T}}{O_{TT}} \times O_{T2}$	O_{1T}
¬ A	$E_{21} = \frac{O_{2T}}{O_{TT}} \times O_{T1}$	$E_{22} = \frac{O_{2T}}{O_{TT}} \times O_{T2}$	O_{2T}
Total	O_{T1}	O_{T2}	O_{TT}

Table 3.3: Contingency table: Expected values

identification, whereas both A und B represent words that might occur within a certain window.

The most popular measure in corpus linguistics to test for association is the log-likelihood ratio Equation (3.10) as presented by Read and Cressie [1988] and reformulated in terms of the contingency tables (Table 3.2, Table 3.3) by Stefanowitsch [2020], presented in Equation (3.10).

$$LLR = G^2 = 2 \sum_{i=1}^{n} O_i \log_e \frac{O_i}{E_i} \tag{3.10}$$

Other popular measures used to be χ^2 statistics or the mutual information. Dunning [1993] showed how these approaches are prone to overvaluing rare events, due to their assumptions of a normal distribution. The LLR on the other hand does not suffer from this assumption. Furthermore, the LLR is partly empirically motivated by its central position in the continuum by Brezina [2018] (Figure 3.2), which is drawn from research experience and not a quantitative measurement. The dimension of 'Exclusivity' can be interpreted similarly to distinctiveness, that is, how much an item as associated only with the foreground corpus, opposed to the background corpus.

For a differentiated presentation and discussion of dispersion measures, please consider Gries [2020]. In the most crude operationalization of dispersion, the *range* describes in how many corpus parts a term appears. Other measures such as Juilland's D [Juilland and Chang-Rodriguez, 1964] or Carrol's D_2 [Carroll, 1970] correct for different sizes of the corpus parts. In the case of Gries [2021], the Kullback-Leibler divergence is calculated between the distribution P of a candidate term over the documents of the corpus and the distribution Q of the document lengths, i.e. $D_{KL}(P \| Q)$.

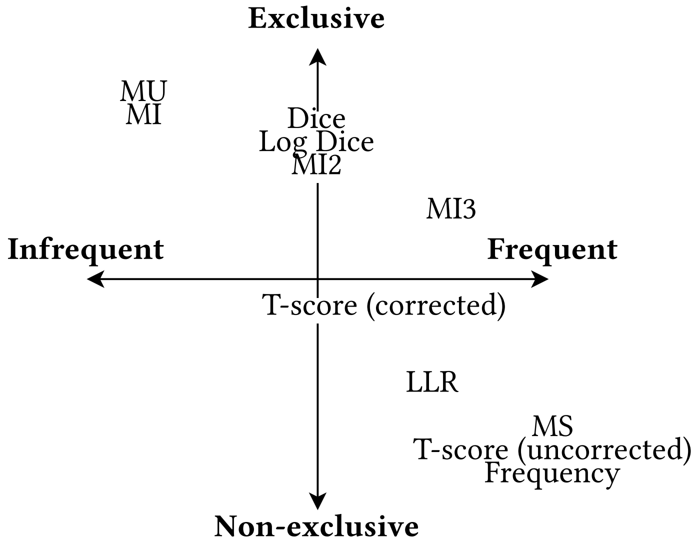


Figure 3.2: Association measures ordered according to Brezina [2018]

3.2.1.3 Phrases are more than consecutive Words: Phraseology

Gries [2008] gives an influential overview on different interpretations of phraseology and the items it is concerned with. He lists six dimensions along which phraseological items ("phraseologism") are usually defined, which are directly quoted below [Gries, 2008, p. 4]:

1. the *nature* of the elements involved in a phraseologism;
2. the *number* of elements involved in a phraseologism;
3. the *number of times* an expression must be observed before it counts as a phraseologism;
4. the *permissible distance* between the elements involved in a phraseologism;
5. the degree of *lexical and syntactic flexibility* of the elements involved;
6. the role that *semantic unity* and *semantic non-compositionality / non-predictability* play in the definition.

As pointed out by Gries [2008], the last point is probably the most important and most interesting in terms of quantitative operationalization. While *semantic unity* and *non-compositionality* are theoretic terms, *non-predictability* can be viewed as a translation of these notions into the domain of information theory and language modeling. Compositionality refers to an aspect of language by which the meaning of large units can be inferred from the meaning of the smaller units that constitute it. Non-compositionality thus describes larger units that have a meaning which cannot be inferred from the parts alone, units such as “it’s not rocket science”. This is tied to the notion of predictability, although it remains underspecified in the chapter by Gries [2008]. It is implicitly assumed that non-compositional phraseologisms are predictable in the sense that certain parts of the units strongly inform the other parts. A (good) language model should be able to predict the phraseologisms with a high probability once parts of it are known.

In terms of KE research, the parameters outlined by Gries [2008] are usually the same: 1) The involved elements are word forms or lemmas appearing in the text, often restricted to certain syntactic patterns as noun phrases, 2) the amount of units ranges from 1 upwards (a distinction from almost all of phraseological research), 3) minimum count filters are pervasive, 4) the units of keyphrases must be contiguous, 5) usually flexibility is of no concern and 6) the *semantic unity* is usually not considered.

The third keyword sense ‘keypatterns’ described by Stubbs [2010], already discussed in the last chapter, is also pertaining to phraseology and keyphrases. The ‘Pattern Grammar’ that studies such keypatterns as fundamental units of meaning is closely related to Construction Grammar [Croft, 2010], CxG for short. Construction Grammar itself goes back to a foundational paper by Fillmore et al. [1988]. They argue for their approach to grammar in part by considering idiomaticity, and posit that idiomaticity goes way beyond non-compositionality. They mention three dimension of idioms:

1. “Encoding vs Decoding Idioms”: Fillmore et al. [1988] explain the difference between encoding and decoding idioms through the lense of second language acquisition. The meaning of decoding idioms can only be known if the idiom itself is known, whereas encoding idioms can be known from previous experience without knowledge of the idiom itself. An example for an encoding idiom is “kick the bucket”, whereas a decoding idiom would be “answer the door” or “bright red” [Fillmore et al., 1988, p. 505]. This distinction is related to (non-)compositionality.

2. “Grammatical vs Extragrammatical Idioms“: Grammatical idioms follow the standard grammar of the language, such as the previous “kick the bucket”. Extragrammatical idioms on the other hand have an “anomalous” structure, such as “all of a sudden”. Fillmore et al. [1988, p. 505] mention how even these extragrammatical idioms follow a structure, but one that is “not made intelligible by knowledge of the familiar rules of the grammar”.

3. ”Substantive vs Formal Idioms“: Gries [2008] touched upon this distinction with his dimension of lexical and syntactical flexibility, although it is not completely congruent. Fillmore et al. [1988, p. 505] consider substantive idioms to be phrases

> with a "(more or less)" fixed lexical inventory such as the aforementioned "kick the bucket". There might be inflectional variation, i.e. "kicked the bucket", but the meaning of the pattern does not translate to a phrase like "kick the water container" or "throw the bucket", or only in a word play sense. On the other hand, "bright red" is a formal idiom, or rather the actualization of the formal idiom "bright COLOUR". The meaning of the idiom translates to "bright blue" or "bright green". Fillmore et al. [1988, p. 506] however go much further than this introductory bigram example for formal idioms, and focus on "general syntactic patterns" such as "The more carefully you do your work, the easier it will get'.

Comparing phraseological considerations from the field of linguistics to the notion of "phrase" in KE research from applied NLP, the latter will show to be rather simplistic. The focus on noun phrases, as will be presented in the next section, is of course a reasonable restrication regarding the otherwise very large search space, and is well grounded in the needs of the application. It was still important to me to mention and introduce these linguistic and phraseological distinctions, as the definitions brought forth by the last 25 years of applied KE research don't really provide a framework equipped for discussing the 'phrase' aspect of keyphrases.

3.2.2 Keyphrase Extraction as an applied NLP task

After already having introduced a brief overview of unsupervised KE in the related works section (Chapter 2), this section will focus mainly on statistical approaches. The role of this secion is partly to discuss the approaches themselves, but also to identify differences to the interests of corpus linguistics. The statistical approaches compile comparatively simple features for candidate items from frequencies and co-occurrences. The oldest automatic KE methodologies are based the statistics of word or phrase frequencies, and modern approaches such as *YAKE!* [Campos et al., 2018] are widely used, due to its domain-agnostic and resourceful nature.

According to Song et al. [2023], the first formulation of keyphrase extraction as an NLP task has been undertaken by Turney [1999]. Although no longer widely cited by papers, the statutes it presented still reverberate today. Turney [1999] derives the task from scientific journals' practice to ask authors for five to fifteen *key words* for their articles. He then introduce the term *keyphrase*, as authors don't necessarily stick to single words, and constitute the research area with this definition:

> "We define a *keyphrase list* [emphasis in the original] as a short list of phrases (typically five to fifteen noun phrases) that capture the main topics discussed in a given document. This paper is concerned with the automatic extraction

> of keyphrases from text. [. . .] We define *automatic keyphrase extraction* as the automatic selection of important, topical phrases from within the body of a document." [Turney, 1999, p. 1]

This definition would prove to be highly influential, as it carves out the tracks research is still on today. The focus it posits on noun phrases is reflected up to the 2020s in the heuristics that identify keyphrase candidates, tying the approaches tightly to the scientific domain with its pronounced nominal style. As far is as I can tell, it is only the latest (influential) research that lifts the Part-of-Speech restrictions posited by other works, either by skipping the candidate selection process entirely, or by employing purely statistical filters (e.g. Campos et al. [2018]; Gu et al. [2021]). Gu et al. [2021] use such a filter to identify maximal contiguous phrases that appear more than once in a document, dubbing them "core phrases". They don't reflect on their motivations, but they don't speak of keyphrases, but of "quality phrases" [Gu et al., 2021], a term I did not find elsewhere in KE research.

All of the big KE survey articles [Hasan and Ng, 2010, 2014; Papagiannopoulou and Tsoumakas, 2020; Song et al., 2023] implicitly show or mention how especially the supervised approaches are geared towards the academic or news domain. This presents itself in features such as citation networks, or ideal document length.

Another idea set by Turney [1999] is the focus on single documents. Not many NLP papers stray from this idea and widen the scope of their keyphrases to encompass more than one document. Tomokiyo and Hurst [2003] do extract keyphrases for a whole corpus, but even when their approach is discussed, this isn't highlighted [Hasan and Ng, 2014; Papagiannopoulou and Tsoumakas, 2020].

Something Turney [1999] doesn't reflect on, but that gets picked up by following research, is the notion of exclusivity or *distinctiveness* [Sönning, 2023]. In many understandings, keywords shouldn't only model a document in a concise way, but also *set it apart from similar documents*. If a paper on a novel machine learning architecture is regarded only in the context of a machine learning paper repository, "machine learning" would make a poor keyphrase. However, if it is a general scientific corpus, it would be a good keyphrase. The probably simplest measure productively used to model this idea is idf in the tfidf-score in Equation (3.11), first proposed by Sparck Jones [1972], although not yet under this name. The number of occurences of term t in document d is denoted as d_t, while the total number of terms in the document is denoted as d. Analoguous, D_t denotes the number of documents where term d is occuring, while D denotes the total number of documents.

$$\mathrm{tfidf}_t = \underbrace{\frac{d_t}{d}}_{\mathrm{tf}} \times \underbrace{\log\frac{D}{D_t}}_{\mathrm{idf}} \tag{3.11}$$

While Sparck Jones [1972] mainly argued for the inverse document frequency with empirical results (and a short pointer to Zipf's law), later research has been looking for a theoretic justification. Due to the similarity of the inverse document frequency with Shannon information, researchers have turned to Information Theory for explanations [Aizawa, 2003]. Following such an interpretation, the idf_t is regarded as the amount of information carried by term t, or rather, the amount of information of t being present in a document. However, such a view runs into problems [Robertson, 2004]. Especially if regarded in the context of IR, wherein idf_t is most often used as a weighting term and combined with term frequency tf_t, different event spaces are being conflated. Other critiques Robertson [2004] enunciates I deem pertaining more to idf_t employed in IR, and not in keyphrase extraction.

When viewing tfidf through the keyness aspects by Sönning [2023] outlined above, tf_t measures a term's discernibility and idf_t the term's distinctiveness.

In the paper introducing topical clustering to KE, Liu et al. [2009] mention three qualities of keyphrases: Keyphrases must be *understandable*, *semantically relevant* to their document and have a high *coverage* of this document. I believe the *understandability* is what leads to the widespread use of syntactic filters. 'Aboutness' or 'topicality' already prefigures a notion of referentiality to a world outside of the text, and concepts and things are most easily referred to, usually with noun phrases. Therefore it is sensible to assume that noun phrases make the best keyphrases, and candidate selection should focus on them. The aspiration of *coverage* sets th cluster-based approaches apart from corpus linguistic research. In corpus linguistic research, key items are regarded on their own, and their qualities are discussed only in their relation to the corpus. In applications however, the key phrases need to be considered next to the other identified key phrases of the document. If three out of five keyphrases describe the same concept, slightly differing in inflection and casing, these three would make a poor keyphrase group, even though they all might be representative of the document.

One of the most popular current KE approaches is *YAKE!* [Campos et al., 2018, 2019; Papagiannopoulou and Tsoumakas, 2020]. After the preprocessing a document, which only includes segmentation, tokenization and stopword flagging, *YAKE!* calculates 5 features for each remaining unigram type, or term, as the authors call it:

1. *Casing* (T_{Case}): The ratio of a term's occurences in uppercase and total occurences is calculated. The higher this value, the more important the term is considered.

2. *Term position* ($T_{Position}$): The distribution of the term over the sentences of the document is considered, with a formula that favors terms occurring near the start of the document.

3. *Normalized term frequency* (TF_{Norm}): The higher a term's frequency, the more it is considered as a key term. Campos et al. [2019] apply an intricate and unusual normalization by dividing the frequency of a term by the mean of all term frequencies plus the standard deviation of this frequency distribution.

4. *Term relatedness to context* (T_{Rel}): This feature implements the assumption that relevant terms only occur in very specific contexts, as opposed to the extreme of stopwords, which occur in all possible contexts. This feature is opposed to the normalized term frequency, as frequent terms have a higher chance of occurring in differing contexts.

5. *Term different sentence* ($TF_{Sentence}$): The number of sentences containing the term compared to the total amount of sentences constitute this last feature. This heuristic models the assumption that important terms occur in more sentences than unimportant terms.

All these features are integrated into a single formula without hyperparameters, seen in Equation (3.12).

$$S(t) = \frac{T_{Rel} \times T_{Position}}{T_{Case} + \frac{TF_{Norm}}{T_{Rel}} + \frac{TF_{Sentence}}{T_{Rel}}} \tag{3.12}$$

A low score corresponds to a term being good keyphrase material. After scoring, keyphrase candidates are formed from the unigrams, by applying a sliding window over the segments of the document. Several empirically validated rules are employed in this phrase formation step, such as not allowing stopwords in boundary positions. The final keyphrase ranking relies on the unigram scores, as well as a new features coming from the frequencies of the keyphrases and their constituting bigrams, to level the field for phrases of different length.

I did not delve into the refined technical details of *YAKE!*, what I wanted to show instead, is how *YAKE!* only relies on the foreground document. None of the features rely on a background or on external knowledge. In the terms of Sönning [2023], *YAKE!* focuses solely on *Discernibility* and *Generality*, although the latter is treated with much more nuance than a term being simply represented in all parts of the corpus. Of course, the descriptive framework by Sönning [2023] focuses on corpus linguistics and as such on linguistic research and text collections, as opposed to application and single documents.

Something that becomes evident by the popularity and quality of *YAKE!* (one of the best approaches in the evaluation of Papagiannopoulou and Tsoumakas [2020]) is how the difference between background-based and background-free approaches is not as clear cut as the framework of Sönning [2023] makes it out to be, at least for the applied setting. Here, the background is, to some extent, encoded in the foreground, especially in the scientific or news domain. Reading relies on background knowledge, certainly the knowledge of the language, but also of the domain. Depending on who an author is addressing, they are presuming a certain extent of background knowledge. The Gricean maxim of quantity for speech acts is of great relevance to this argument: "Contributions [need to be] as informative as is required" but not "more informative than is required [. . .] (for the current purposes of the exchange)" [Grice, 1975, p. 45]. The very foreground text itself already informs about what the background is, as the information (not) conveyed positions it at a very specific location in the existing discourse.

This is a rather htheoretic way to say that texts are about things a reader does not yet know in this form, but that they assume the reader to know some other things to understand them. I believe it is exactly this principle that allows the foreground-focused approach of YAKE! to be as successful as it is, together with all the nuanced text linguistic heuristics it employs.

3.3 Language Models

Language models are models that are historically understood to assign probabilities to sequences of language. The most basic language models are so called n-gram models, where the probability of a word is determined by its n-1 preceding words. This leads us to a terminology where a unigram language model ($n = 1$) uses no preceding information to ascribe a probability to a word w_k, whereas a bigram model ($n = 2$) is informed by a single preceding word w_{k-1}. A language model can be inferred from a corpus via the *maximum likelihood estimation*, i.e. by obtaining n-gram frequencies from a corpus and normalizing them [Jurafsky and Martin, 2023]. If deriving a bigram model from a corpus where word A appears 2 times followed by B and 8 times followed by other words, then the probability $p(B \mid A)$ is 0.2. In practice, smoothing procedures are applied.

3.3.1 Perplexity

Causal language models, i.e. language models that only informed by previous context, are frequently intrinsically evaluated using *perplexity* (Equation (3.13a)), which is directly tied to the entropy [Jurafsky and Martin, 2023] (Equation (3.13b)).

$$
\begin{aligned}
perplexity(W) &= P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} && \text{(3.13a)} \\
&= 2^{H(W)} && \text{(3.13b)}
\end{aligned}
$$

W denotes a sequence of words, and the entropy $H(W)$ is calculated based on the distributions of the language model. The whole sequence is considered a single event. For a detailed discussion of how this can be achieved, please consider Jurafsky and Martin [2023, p. 52ff]. In a nutshell, the entropy here is now actually the *entropy rate*, a „per-word-entropy“ [Jurafsky and Martin, 2023, p. 53], and it assumes that language is a a stationary ergodic process, meaning that the distributions that generate the words never change over time.[5]

HuggingFace, while a not a scientific source, gives Equation (3.14) for the perplexity (PPL), in the context of evaluating causal language models [HuggingFace]. In comparison to n-gram language models, the conditionals for token x are all of the preceding tokens $x_{<i}$.

$$
\operatorname{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t} \log p_\theta\left(x_i \mid x_{<i}\right)\right\} \tag{3.14}
$$

Masked language modeling views language fundamentally different then the previously discussed autoregressive language models. MLMs are trained by filling in gaps in sentences, where the whole rest of the sequence can inform the models decision. Perplexity and (cross) entropy over the whole sequence can therefore not be computed in the same way as for causal language models. Salazar et al. [2020], building upon [Wang and Cho, 2019] propose the use of pseudo-log-likelihood (PLL) scores which they define as in Equation (3.15)

$$
\operatorname{PLL}(W) := \sum_{t=1}^{|W|} \log P_{\mathrm{MLM}}\left(w_t \mid W_{\backslash t}\right) \tag{3.15}
$$

I want to highlight a few characteristics and differences to the definition of perplexity and entropy in the case of causal language modeling. Firstly, the probability of w_i is not conditioned on the tokens $w_{<i}$ preceding it, but on the whole sentence except this one token $W_{\backslash t}$. Furthermore, they don’t add transformations such as turning the term

[5] Some sources name the entropy to the base 2 as being equivalent to the perplexity, while others name the cross-entropy. In fact, due to the assumptions of the Shannon-McMillan-Breiman-theorem, the reformulation of the two become equal [Jurafsky and Martin, 2023, p. 53f].

negative or dividing by the number of words, which would make a resemblance to the (cross) entropy under the Shannon-McMillan-Breiman theorem.

Salazar et al. [2020] define the pseudo-perplexity (PPPL) as in Equation (3.16):

$$\mathrm{PPPL}(\mathbb{W}) := \exp\left(-\frac{1}{N}\sum_{\boldsymbol{W}\in\mathbb{W}} \mathrm{PLL}(\boldsymbol{W})\right) \tag{3.16}$$

Please note how $\mathbb{W}$ is the set of all sentences in the corpus, and not all words. The pseudo-log-likelihood has been scrutinized by Kauf and Ivanova [2023], who show how masks that are placed on subword tokens which are part of multi-token words distort the score. In experiments they compare the definition proposed by Salazar et al. [2020] with two other strategies. In one formulation, all subword tokens of a word are masked when determining the probability of one of those subwords. The other formulation only masks the subwords of the same word if they are to the right of the token in question. While the original formulation by Salazar et al. [2020] attributed lower and lower negative PLL scores (i.e. lower surprisal) the longer the sentences are, the negative PLL calculated by the two word-aware strategies increases with sentence length, following the same (desired) trend as with the log-likelihood of causal language models. This finding supports the hypothesis that subword tokens are strongly informed by subtokens of the same word.

3.3.2 Adapters, *X-MOD* and *SwissBERT*

This chapter presents architectures and innovations that lead up to the model used in this thesis.

BERT [Devlin et al., 2019] was the first widely used pretrained language model that broke with the causal or autoregressive tradition. *ELMo* [Peters et al., 2018] already combined features for token representations when the sentence was regarded from left-to-right and from right-to-left, but was not yet „deeply bidirectional“ [Devlin et al., 2019]. *BERT* achieves this deep bidirectionality by introducing an alternative language modeling objective that was no longer framed as complete-the-sentence, but fill-in-the-gap: *Masked language modeling*, inspired by the "Cloze Procedure" for readability estimation [Taylor, 1953].

BERT consists of a transformer encoder [Vaswani et al., 2017], which during pretraining is tasked to predict masked tokens from two input sentences.[6] The input sentence have

[6] Effectively, while 15 % of tokens are randomly chosen to be masked, only 80 % of these chosen tokens are replaced with the `[MASK]` token. 10 % are changed to another token and 10 % are left as is. Devlin et al.

a 50 % chance of appearing consecutively in the corpus. As a second task, the model needs to predict whether the sentence appear together or not (next sentence prediction). When finetuning *BERT*, the last prediction layer used for pretraining is removed and a task specific layer or architecture is added on top of the pretrained encoder block.

Liu et al. [2019] investigate *BERT* and its hyperparameters and conclude by presenting *RoBERTa*, an improved iteration upon the original *BERT*. They now only pretrain on masked language modeling, and they use a dynamic masking strategy, meaning that the same sentence can appear with different masks across epochs. Furthermore, Liu et al. [2019] discovered how *RoBERTa* could achieve performance gains by training for much longer than *BERT* originally did.

Lample and Conneau [2019] introduced *XLM*, an architecture for cross-lingual language modeling. They evaluate causal language modeling, masked language modeling as well as a new objective for translation language modeling. For translation language modeling, they concatenate parallel translations and task the model to infer masked tokens in both sentences. A single *SentencePiece* [Kudo and Richardson, 2018] tokenizer is trained on all languages, and special language embeddings are concatenated to the token embeddings. Based on the *BERT* architecture, they report a significant improvement when using translation language modeling in addition to masked language modeling for crosslingual tasks. A year later, Conneau et al. [2020] present *XLM-R*, an improvement upon *XLM*, partially based on the findings on scaling of the *RoBERTa* paper [Liu et al., 2019]. Furthermore, they abandon the language embeddings, motivated by phenomena such as code-switching, where single sentences can no longer be attributed to a single language.

Since the introduction of the transformer architecture, the notion of improvements by adding more data, more parameters and more training, effectively by adding more resources, has become a truism. While some papers and innovations focused on more scaling, others focused on more efficient training and architectures. In the same year as *RoBERTa*, Houlsby et al. [2019] presented the concept and the efficiency of adding small layers into existing pretrained language models and only finetuning on those: Enter adapters. In the original paper, adapters are so called bottleneck layers. They project the internal state of a transformer block to a lower dimensionality, apply non-linearity, and project the compressed state back to the original size. Houlsby et al. [2019] show how by only adding and finetuning 3% of the parameters of *BERT*, they are only 0.4%pt worse than the fully finetuned *BERT* on the GLUE [Wang et al., 2019] benchmark. This presents an interesting opportunity for multitask settings. For different tasks, different

[2019] motivate this to remedy the discrepancy between pretraining and finetuning, as the `[MASK]` token only appears during pretraining.

sets of adapters can be trained, for a fraction of the cost as if one was training the full model for every single task. This removes the danger of forgetting certain tasks, but also removes the opportunity to let tasks inform each other.

This is addressed by the *X-MOD* architecture [Pfeiffer et al., 2022], although not directly for multitask learning, but for multilinguality. Instead of finetuning adapter layers in pretrained language models, *X-MOD* already uses adapters during pretraining. For each modeled language, specific adapter blocks are inserted into the larger model. The shared parts of the model are still subjected to updates coming from all languages, while the adapters can specialize for their specific language and maintain their knowledge thereof. The architecture is based on *BERT* and *XLM-R*, but the only training objective is a masked language modeling objective. Compared to a model where all parameters are shared across all languages, the performance of *X-MOD* remains high even when more languages are added. Pfeiffer et al. [2022] arrive at the best scores with 60 languages trained simultaneously, and conjecture that the performance drops after that due to decreasing availability of data for the languages they added after the initial 60.

The previous architectures and models presented innovations and techniques that lead up to the model used in this thesis: *SwissBERT* [Vamvas et al., 2023]. *SwissBERT* is motivated by the unique linguistic landscape of Switzerland. German, French, Italian and Romansh are all recognized national languages, and up until *SwissBERT*, have not yet been integrated together into a single multilingual model. While different variants are evaluated, the model published on HuggingFace is based on *X-MOD* with four different language adapters for each national language. Additionally, a new *SentencePiece* tokenizer was trained on the data in all four languages, resulting in a vocabulary with 50'000 tokens. While no language aligning training objectives are employed (as compared to *XLM-R*), Vamvas et al. [2023] conjecture that the overlap of extralinguistic references to the same entities will implicitly guide the model to multilingual representations.

The data used to train *SwissBERT* consists of Swiss news articles published up until the end of 2022, accessed via the the services provided by Swissdox@LiRI. For a discussion of the data please consider Section 4.1. *SwissBERT* is trained using the hyperparameters defined for *X-MOD*. However, instead of training on sentences, *SwissBERT* chunks the articles into spans of 512 tokens and uses them as training samples. Vamvas et al. [2023] motivate this in terms of training efficiency.

3.4 Keyphrase Extraction according to Tomokiyo and Hurst [2003]

The KE approach developed by Tomokiyo and Hurst [2003] stands alone in the broad methodological landscape on keyphrases. Their method is unsupervised and conceptually simple, without the need for hyperparameter tuning. The key insight of their method is that the differences between language models are meaningful and can be harnessed through the methods of information theory.

They motivate their approach not by summarization or bibliographic keywording, but by exploration. Their approach extracts keywords for *sets of documents*, and not for a single document, setting it apart from the approaches within the scope of KE as defined by Turney [1999]. It also makes it ideal for the exploration undertaken in this thesis, with exploratory, data-driven corpus linguistics as an application in mind.

Tomokiyo and Hurst [2003] name two different defining qualities of keyphrases: *Phraseness* and *informativeness*. Phraseness refers to the extent in which a sequence of words or tokens can be considered a phrase. While they mention use cases in which syntactic constraints guide this notion, they focus on “collocation or cohesion of consecutive words” [Tomokiyo and Hurst, 2003, p. 1], once again being ideally suited for data-driven corpus linguistic research. Informativeness however is meant to describe “how well a phrase captures or illustrates the key ideas in a set of documents” [Tomokiyo and Hurst, 2003, p. 1]. They directly tie this definition to the concepts of foreground and background, and point out how defining characteristics of the foreground differ when comparing to a different background. The target document set is therefore dubbed “foreground corpus” and the reference corpus is dubbed “background corpus”.

Sequences of words may have a high phraseness and low informativeness, or vice versa. Tomokiyo and Hurst [2003] want only sequences that are both strong in phraseness and informativeness. As a negative example they mention “in spite of”, a phraseme with high phraseness but low informativeness. For the context of my thesis and interests, I oppose this example, and argue that informativeness *only* depends on the foreground and its relation to the background, and not on a researcher’s notion of referential semantic content. Imagine two corpora of two different public speakers, where one argues for their viewpoint “in spite of” arguments they acknowledge, whereas the other speaker does not even acknowledge any counter points. An exploratory analysis of these corpora would be fruitful if it yielded “in spite of” as a keyphrase for the first speaker.

As a baseline, Tomokiyo and Hurst [2003] use a slightly reformulated binomial log-likelihood ratio test (BLRT) score proposed by Dunning [1993]. The BLRT compares two hypotheses, in the numerator the hypothesis that two events 1 and 2, e.g. the occurrence of token 1 and 2, are drawn from different distributions and in the denominator whether they come from the same distribution.

$$BLRT_{score} = 2\log\frac{L(p_1, k_1, n_1)L(p_2, k_2, n_2)}{L(p, k_1, n_1)L(p, k_2, n_2)} \tag{3.17}$$

$L(p, k, n)$ is the likelihood function assuming a binomial distribution.

$$L(p, k, n) = p^k(1-p)^{n-k} \tag{3.18}$$

The probability p follows from the maximum likelihood estimation, so that $p_i = \frac{k_i}{n_i}$ and $p = \frac{k_1+k_2}{n_1+n_2}$. For informativeness, n is the total number of tokens in either the foreground or background corpus, k is the number of occurrences in either the foreground or background corpus. For phraseness, k_1 is the number of two token x, y cooccurring as a phrase, with n_1 being the count of token x (modeling $p(y|x)$) k_2 on the other hand is the count of any other token appearing in front of y, with n_2 being the count of every other token than x. For phraseness, this assesses whether the presence of x has an influence on the probability of y or not.

A phrase's final score is the parameterized composition of two BLRT scores, one for phraseness, and one for informativeness. For a bigram $(w1, w2)$, the phraseness is computed by comparing the distributions of w_2 given or not given w_1. The phraseness of a word w_n is calculated comparing the distributions of said word in the foreground and background corpus.

Tomokiyo and Hurst [2003] critique their baseline for two things: Firstly, weights need to be trained to combine the two scores, and they desire an approach without parameters. Secondly, the BLRT assigns high scores whenever the distributions are unequal. However, if it finds a difference, it is unclear which of the events are positively or negatively associated. In the case of informativeness, words that are informative of either corpus, foreground or background, get a high score.

Given these critiques, a measure is presented based on the Kullback-Leibler divergence of different language models. N-gram models of different orders are trained for both the background and foreground corpus, with the notation $LM^N_{\{bg,fg\}}$ for a specific model. Phraseness can be measured when comparing the Unigram model LM^1_{fg} to models of a higher order LM^N_{fg}. Informativeness on the other hand is obtained comparing a foreground model LM^N_{fg} to a background model LM^N_{bg}. Instead of BLRT they propose the measure of the pointwise Kullback-Leibler divergence, shown in Equation (3.19).

$$pD_{KL}\big(p(w)\|q(w)\big) = p(w)\log_2 \frac{p(w)}{q(w)} \tag{3.19}$$

Remember how the Kullback-Leibler divergence describes the average excess amount of bits when encoding distribution P based on distribution Q. The pointwise version measures how a single event w contributes to this average. In contrast to the original D_{KL}, pD_{KL} can become negative, in the case where $q(w)$ is larger than $p(w)$.

For informativeness, $p(w)$ comes from the probability assigned to a word by LM^1_{fg} and $q(w)$ comes from LM^1_{bg}. For phraseness, $p(w)$ comes from LM^N_{fg}, whereas $q(w)$ comes from LM^1_{fg}. Tomokiyo and Hurst [2003] note how this definition is closely related to pointwise mutual information when calculating the phraseness of a bigram x, y.

$$pD_{KL}\big(p(x,y)\|p(x)p(y)\big) = p(x,y) \times \underbrace{\log_2 \frac{p(x,y)}{p(x)p(y)}}_{\text{pointwise mutual information}} \tag{3.20}$$

Outside of information theory and in practical terms, the additional term $p(x, y)$ can be considered a weighting factor, as pointwise mutual information as a collocation measure is often critiqued for overrepresenting infrequent terms (consider Figure 3.2 in Section 3.2.1.2). The final score for ranking keyphrases results from simply adding the phraseness and informativeness score.

Given these theoretical foundations, Tomokiyo and Hurst [2003] give the following approach to extract a list of key bigrams. In a first step, four n-gram language models are compiled, for the foreground and background corpus and for unigrams and bigrams. Stopwords are already filtered and not included in the language model. The phraseness of bigram x, y is calculated as $pD_{KL}\big(p_{fg}(x,y)\|p_{fg}(x)p_{fg}(y)\big)$ (Equation (3.20)), with the probabilities being derived from LM^1_{fg} and LM^2_{fg}. I will refer to $p(x, y)$ or $p(t_1, \ldots, t_n)$ in the general case as **phrase-dependent probability**. In turn, $p(t_1)p(t_2)\ldots p(t_n)$ will be dubbed **phrase-independent probability**.

The informativeness is calculated as $pD_{KL}\big(p_{fg}(x,y)\|p_{bg}(x,y)\big)$, with Katz smoothing applied to the background corpus. In essence, Katz smoothing reserves an amount of the probability mass for unseen events, and distributes it according to the probabilities of lower-order models [Chen and Goodman, 1999].

Tomokiyo and Hurst [2003] extend their approach to $n > 2$-grams by bootstrapping from the identified key bigrams. They employ an adapted APriori algorithm [Agrawal and Srikant, 1994] to identify overlapping bigrams and combine them to higher-order phrases. The combined phrases are filtered based on a minimum occurence of 5 and a

Rank	*current-films*	*hybrid-cars*
1	minority report	civic hybrid
2	box office	honda civic hybrid
3	scooby doo	toyota prius
4	sixth sense	electric motor
5	national guard	honda civic
6	bourne identity	fuel cell
7	air national guard	hybrid cars
8	united states	honda insight
9	phantom menace	battery pack
10	special effects	sports cars
11	hotel room	civic si
12	comic book	hybrid car
13	blair witch project	civic lx
14	short story	focus fcv
15	real life	fuel cells
16	jude law	hybrid vehicles
17	iron giant	tour de sol
18	bin laden	years ago
19	black people	daily driver
20	opening weekend	jetta tdi

Table 3.4: Keyphrases reported by Tomokiyo and Hurst [2003]

syntactic filter to only allow noun phrases. A threshold for the key bigrams used for the extension (e.g. up to rank 1000 or a phraseness > 0) is not reported. The extended phrases are ranked as well, and added to the ranking. It is very important to note that the final scores of phrases with different lengths are of a similar magnitude, and bigrams and trigrams appear interspersed and don't need to be reweighted. From the original paper, it is unclear how this behaviour continues for $n > 3$, as they themselves mention how most $n > 3$-grams are filtered out due to the minimum frequency. sref

They evaluate their approach on the 20 newsgroups data set, which contains thematic discussions from 1993, as a background corpus, and 20,000 messages from a *current-films* newsgroup as the foreground. In addition, a list for the subcorpus *hybrid-cars* is also reported, but it remains unclear what was actually used as a foreground and background corpus for this example.

The approach was not evaluated programmatically, but via the manual inspection of the extracted keyphrases as reported in Table 3.4. A thematic corpus lends itself nicely for manual evaluation, because the theme already dictates what good keyphrases should revolve around. Considering this, the lists appear to fulfill the task nicely. Movies that appeared in 2002 are highly ranked for *current-films*, also featuring trigrams such as “bair witch project”. The same is true for *hybrid-cars*, featuring car models and technical parts used in alternative fuel vehicles. Weaknesses show where components of the trigrams appear as well as bigrams, prominently on the first two ranks for *hybrid-cars*. Furthermore, from an application-based perspective, the inclusion of phrases that only differ in inflection such “hybrid cars” and “hybrid car” is not ideal. These problems could both easily be mitigated by pre- and postprocessing.

3.5 Evaluation

Passonneau and Mani [2014] implicitly define evaluation as a process that determines how well a system works. They define four dimensions of evaluation outlined below:

1. *Intrinsic evaluation* vs *extrinsic evaluation*;
2. *Stand-alone application* vs *component application*;
3. *Manual assessments* vs *automatic metrics*;
4. *Laboratory evaluation* vs *real-world context*;

I want to highlight the contrast of *manual assessments* and *automatic metrics*. Key concepts to discuss the differences are *reliability* and *validity* [Passonneau and Mani, 2014]. Reliability refers to how well an evaluation method reproduces the same results across different settings, whereas validity refers to how well a method actually measures what should be measured. Manual assessment usually has a high validity, while the reliability is difficult to achieve. Frameworks that help with reliability have been developed, Passonneau and Mani [2014] mention the EAGLES standard, checklists, guidelines and rater agreement. All these frameworks however add to how resource-intensive manual assessment is. Automatic metrics on the other hand are cheap to deploy and easily to reproduce, but do not necessarily correlate well with human judgment.

A dimension not discussed by Passonneau and Mani [2014] is one I will dub *reference-free* vs *reference-based*, borrowing from the evaluation of machine translation methods. Reference-based evaluation relies on a gold standard, whereas reference-free evaluation either comes from human judgment or by defining desired metrics for the output, such as a smooth label distribution for classifiers or similar cluster sizes for clustering tasks.

3.5.1 Evaluating Keyphrases with a Gold Standard

Most research on KE uses manually compiled gold keyphrases to evaluate their models and algorithms. Because the scope of their understanding of keyphrases doesn't usually extend over a single document, keyphrases are not far from reach. Either because a human annotator can read the document in short time, or because keyphrases already exist in the metadata, e.g. for scientific papers or news articles (e.g. Boudin [2013], Caragea et al. [2014] or Gallina et al. [2019]).

The existence of a gold standard allows for an automatic evaluation. Zesch and Gurevych [2009] propose R-precision (R-p) to evaluate KE system. R-p is calculated for a single document, and it presupposes that the amount of extracted and gold keyphrases is equal. In practice, the gold standard therefore determines the cutoff for the ranking of the extracted keyphrases. R-p is then defined as the ratio between the amount of retrieved ("matching" keyphrases) and the amount of gold keyphrases. If the cutoff is determined by the amount of keyphrases in the gold standard, precision and recall will be the same (consider equations 3.21 and 3.22). In contrast to previous evaluation regimes, Zesch and Gurevych [2009] also introduce fuzzy matching for determining whether an extracted keyphrase matches one from the gold standard.

In the recent survey by Papagiannopoulou and Tsoumakas [2020], precision, recall and F_1 are evaluated at two different cutoff-points, 10 and 20 keyphrases. They combine all their evaluation metrics with different matching strategies, from exact to partial matching.

$$\text{precision} = \frac{\text{number of correctly matched}}{\text{total number of extracted}} \tag{3.21}$$

$$\text{recall} = \frac{\text{number of correctly matched}}{\text{total number in gold standard}} \tag{3.22}$$

$$\mathrm{F}_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3.23}$$

Additionally, they present different measures that also consider the ranking order, and choose the Mean Average Precision (MAP, in Equation (3.25)) to conduct their comparative study. For the average precision (AP, Equation (3.24)), $|L|$ is the number of keyphrases in the ranking, $P(r)$ is the precision when considering only the first r items of the list, $rel(r)$ is 1 if the r^{th} item of the list is in the gold standard or 0 otherwise. $|L_R|$ is the averaging term, denoting the number of correct matchings. The MAP is then obtained by averaging the AP over all documents.

$$\mathrm{AP} = \frac{1}{|L_R|} \times \sum_{r=1}^{|L|} P(r) \times rel(r) \tag{3.24}$$

$$\mathrm{MAP} = \frac{1}{n} \times \sum_{i=1}^{n} AP_i \tag{3.25}$$

In corpus linguistic research, evaluation of methods by means of a gold standard is undertaken as well. However, they rarely focus on only the top 10 or 20 extracted keywords or -phrases, but are more interested in the behavior of a measure or procedure as a whole. For example, Sönning [2023] evaluates different measures to identify "academic key verbs" by using a list of verbs of a previous corpus study by Paquot [2010]. Paquot [2010] herself used a "(semi-)automatic extraction procedure", based on the log-likelihood ratio and a dispersion measure. While Paquot [2010] uses a very informed approach, the inherent problem of using an automatically extracted gold standard to evaluate other models remains for the study of Sönning [2023]. Nonetheless, he focuses on criteria he dubs on *coverage* and *reliability.* The *coverage*, not to be confused with the topical coverage of Liu et al. [2009], is essentially *R-p* proposed by Zesch and Gurevych [2009] and discussed above. Sönning [2023] splits a large scientific corpus into 100 subcorpora, and separately calculates keywords for each subcorpus separately. He then analyzes how the coverage behaves when comparing across all subcorpora. The reliability is reference-less and does not depend on the gold standard, but compares the ranking orders over all subcorpora, quantified with a reliability coefficient.

3.5.2 Evaluating Keyphrases without a Gold Standard

Few systematic reviews of different keyness approaches exist, and when they do, they often recur to previously compiled lists as gold standard (e.g. Paquot [2010]. However, these analyses are then accompanied by reference-less analyses. Previous corpus linguistic research realized how different measures differ strongly when considering the frequency of the items in question, and therefore, the behaviour of measures over different frequencies are reported (Gries [2021]).

In a similar vein, Brezina [2018] graphically compares different association measures along two dimensions (Figure 3.2, already discussed in Section 3.2.1.2). In the case of the approach by Brezina [2018], it is not the results of the methods being evaluated, but rather the methods themselves being discussed. Depending on the research interest, a method can already be discarded that way.

Another method of not necessarily evaluating but discussing the methods behind keyphrases comes from the comparison of different methods on the same corpus, and analyzing where they differ. Evert et al. [2018] follow this approach when comparing three different approaches, the LLR being one of them. They compare the overlap of the top 200 keywords identified by the approaches and how these top 200 lists distribute over the frequencies in the foreground corpus (not before enforcing a minimum frequency of 5 in the reference corpus). Furthermore, they use a previous qualitative study on the same corpus to classify the identified keywords into "key categories" discovered in said study and use it to calculate precision. A keyword is considered a true positive if it can be ascribed to a category, and a false positive otherwise.

A final way of evaluating or rather discussing metrics is strictly rationalistic, without any recursion to empirical results. When examining the theoretical assumptions of formulae and algorithms, researchers can highlight how some presuppositions are undesirable. Dunning [1993] proceeds in such a way when he criticizes χ^2 or z-score tests for the analysis of textual data, because their assumption of normality does not hold for rare events, i.e. most words when considering the Zipfian distribution. He then moves on to present another metric, the log-likelihood ratio, which does not suffer from this assumption. Also Gries [2021] arrives at the Kullback-Leibler divergence as a keyness metric by scrutinizing the theoretical assumptions of previous methods. Another discussion method, employed for example by Gries [2020], uses dummy data to exemplify and analyze properties of different methods. Of course, these theoretical discussions are always embedded in a cycle of experimentation and empirical validation. Nonetheless, I want to stress the importance of theoretical scrutiny next to empirical evaluation.

3.5.3 Evaluation in this thesis

This thesis is concerned with exploring, evaluating and understanding a corpus-level KE approach. A gold-standard is not attainable, because the corpus is both too big and too recent for qualitative researchers to have created categories like the ones Evert et al. [2018] profited from. Therefore, I am left with evaluation or discussion techniques that do not rely on a gold standard.

The approach I adapt from Tomokiyo and Hurst [2003] will be be discussed both from a theoretic and an empiric perspective, by comparing its formulae and results to different variants of itself and the corpus linguistic LLR. In the preceding chapters, I have collected and introduced several qualities and requirements towards keyphrases, and this terminological toolbox will be used too describe and compare the results in as much detail as possible.

4 Data

4.1 Swissdox@LiRI

Swissdox@LiRI[1] is an initiative providing a programmatic interface to the Schweizer Mediendatenbank (SMD). The SMD is a private company owned by the three biggest swiss media outlets Ringier, Tamedia and the public broadcasting company SRG SSR. Selected news papers and magazines of other publishing companies are indexed and available as well. It is furthermore important to note that the SMD is not media history carved in stone. It may be subject to censorship and revision once a judicial body has determined a violation of personal rights in certain news articles, or for any other reason.

The interface of Swissdox provides a simple query system and delivers the retrieved data in a tabular format. The columns contain metadata such as the date of publishing or the news outlet, as well as the text of the article with basic SGML markup. There exist inconsistencies in how the unstructured data (i.e. the text) is treated. Three fields are reserved for the content of the news article: HEAD, SUBHEAD and CONTENT. The HEAD is the title of an article, while SUBHEAD is a very short introduction to the article, rarely longer than two sentences. The SUBHEAD is an optional field, but often in its absence, the CONTENT begins with an `<ld>` element, short for "lead", the industry term for the aforementioned SUBHEAD. This difference is likely due to different digitization or publishing procedures and poses no problem to the presented research or the research it builds upon.

When regarding the text linguistic dimension of the data, two things are important to note: First, on a structural level, the (online) news domain uses the semantics of layouting. While seemingly trivial and true for almost all text types, this presents a challenge when automatically processing the text. As the layout in its technical representation is mainly conveyed through markup elements, and these elements are stripped in the cleaning process, information is lost. This is especially troublesome in the case of sentence boundary identification. Usually, titles of sections and subsections don't form complete grammatical sentences, but can be parsed as such by human readers due to typographic

[1] `https://swissdox.linguistik.uzh.ch/`

contrast and linebreaks. The same is true for lists and tables, prevalent in live feeds and sports reporting. If preprocesing doesn't account for this, we need to assume both added noise in the case of training language models as well as erroneous sentence boundary identification. The size if the impact however may be negligible. Secondly, on a semantic and discursive level, media articles are most often concerned with current world affairs and therefore strongly temporally situated. Topics, and the words and phrases associated with these topics, can be expected to disperse unevenly along the temporal axis. This insight, however trivial, will prove vital when discussing the notion of keyness in the case of a general news corpus.

A special quirk of the data lies in the amount and types of duplicates. Due to the oligopolization of the swiss media market, articles are shared among news papers and sites that belong to the same publisher. For the year of 2023, the **?** determined that every fourth news article is shared among multiple outlets. This is a current phenomenon, as in 2017 the share of shared articles was only at 10 percent [**?**, p. 162]. Another type of duplicate is on the quasi-sentence-level. For one, this is noticable in leads and titles of articles that are continuously updated and re-published, with recurring phrases as "Die aktuellen Corona-Zahlen" or "Krieg in der Ukraine - die Übersicht". Other phrases such as source references for short news (e.g. "(Quelle: SDA)") or photographs are highly formalized and strongly deviate from the distribution of other parts of the language used.

4.2 Descriptive Statistics

To describe the data that will be used as a background corpus, I need to refer to Vamvas et al. [2023], because the full dataset *SwissBERT* was pretrained on is noo longer available to me. According to Vamvas et al. [2023], the pretraining data contains 21,595,088 articles with more than 14 Billion subword tokens, with almost 18 Million articles and 12.6 billion subword tokens in German. It is the German adapter (`de_CH`) that will be used in my experiments. The publications range from 1911 to the end of 2022, the large majority has been published after the early nineties.

For evaluation purposes, I sampled the background corpus by downloading German articles from 1992 to 2022 via Swissdox@LiRI, sampling four different weeks for every year, therefore with a downsampling factor of $52/4 = 12$. This sampled background corpus contains 1,520,655 articles, approximately 12 times less articles than the number reported for the *SwissBERT* pretraining corpus. The obtained sample is then again downsampled with a factor of 10, which resulted in a background corpus of around 150,000 articles.

The foreground corpus consists of all German articles that are available on Swissdox@LiRI for the month of January 2023. This amounts to 53,156 articles that have been used for finetuning. However, because some of the articles contain duplicates that are not filtered out by the preprocessing suite of *SwissBERT*, and the approaches have proved to be sensitive to duplicate text, I created a second foreground corpus for without duplicate articles for inference, amounting to 47,864 articles. Roughly 10 percent of the articles were filtered out. I assume I did not arrive at the 25 percent reported by **?** because I kept one article per duplicate group and because they only considered newspapers belonging to two major publishing companies, whereas Swissdox@LiRI and the SMD contain other outlets as well.

5 Experiments

This thesis presents two main groups of experiments. The first one, presented and discussed in this chapter, is concerned with the reproduction and adaptation of the KE approach by Tomokiyo and Hurst [2003], which was discussed in Section 3.4. The second experiment group takes the insights of the first and distances itself from the original paper, both in its approach and scope.

5.1 Adaptation of Tomokiyo and Hurst [2003]

5.1.1 Methodology

Tomokiyo and Hurst [2003] build their approach around the probabilities assigned to words by different language models. They explicitly mention how their method is open to all kinds of language models, and not restricted to n-gram models. This thesis will put this claim to the test, 20 years later, by extracting the probabilities from one of the newest generations of language models: *SwissBERT*. The transfer of the original approach to this architecture is not straight forward, and needs to be argued for. The main hurdle is the difference between the original causal n-gram language modeling and the masked language modeling of *SwissBERT*.

5.1.1.1 Foreground and Background Model

Due to the resource-intensive nature of training modern language models as *X-MOD*, it is not feasible to fully train two models from scratch. However, the idea of pretraining lets us still arrive at a setup conceptually similar to Tomokiyo and Hurst [2003]. The main idea behind a foreground and a background model is that they model a certain corpus of text as best as they can, i.e. model the corpus with the lowest perplexity possible. When finetuning a language model, it is biased away from the original distribution and towards the distribution of the new data. This means that *SwissBERT* can be finetuned on new data, which it will then model with a lower perplexity than the original *SwissBERT*.

Because I don't use a different objective than during pretraining, i.e. MLM, one could also speak of continuing the pretraining. I will adhere to the terminology of "finetuning", as it is my intention to bias the model towards my new data and away from the pretraining data, and not towards some middle ground. Using the data preprocessing of the original *SwissBERT* [Vamvas et al., 2023], a dataset is created from all SwissDox articles of the January of 2023. This is directly motivated by the way we want to the foreground model to be biased. The differences between foreground and background model should be solely based on the thematic content of the texts, and not any distributional artifacts introduced by preprocessing.

The preprocessing consists of removing markup embedded in the retrieved document and replacing it by a special token `</s>` [Vamvas et al., 2023]. During pretraining, some metadata as the date of publication is prepended to the article, separated by the same token. The texts are not split into sentences, but into contiguous token blocks of 512 tokens. The model with the German adapter (`de_CH` is finetuned on the same MLM task with a chance of 0.15 for masks, a batch size of 4 and a learning rate of $1e-5$ for 3 epochs.

5.1.1.2 Pointwise Kullback-Leibler Divergence in the Case of MLM

The challenge of adapting Tomokiyo and Hurst [2003] to an MLM arises especially for phraseness. In the original paper, two hypotheses are compared: Empirical occurence (phrase-dependent) and expected occurence when assuming independence between the phrase parts (phrase-independent). The accompanying probabilities can easily be extracted from the n-gram language model, as it essentially consists of these probabilities. The same thing cannot be said for MLMs. Single masks are maximally informed by their context, and masking the whole sentence to get a unigram probability would defeat the purposes of the model itself.

Therefore I will take a step away from the formulations by Tomokiyo and Hurst [2003] and examine the abstract ideas behind them. For the case of phraseness, empirical and chance distribution of a phrase are compared with the pointwise Kullback-Leibler divergence, essentially arriving at a weighted Mutual Information of the phrase parts. The problem can be rephrased as how strongly the parts of a potential phrase inform each other, with a weighting term based on the phrase's empirical, phrase-dependent probability (refer to Equation (3.20) in Section 3.4). In the case of the aforementioned n-gram-language models, the weight is therefore directly based on frequency – rare phrases get a low weight, common phrases get a high weight.

	Tomokiyo and Hurst [2003]	Ideal Adaptation	Simplified Adaptation
Phraseness			
phrase-dependent	$p_{\text{fg}}(t_1, t_2, \dots, t_n)$	$\prod_{i=1}^{n} p_{\text{fg}}(t_i \vert T \setminus t_i, C)$	$\prod_{i=1}^{n} \frac{1}{2} \sum_{j\in\{1,-1\}} p_{\text{fg}}(t_i \vert T \setminus \{t_{i+j}\}, C)$
phrase-independent	$\prod_{i=1}^{n} p_{\text{fg}}(t_i)$	$\prod_{i=1}^{n} p_{\text{fg}}(t_i \vert C)$	$\prod_{i=1}^{n} p_{\text{fg}}(t_i \vert C)$
Keyness			
foreground	$p_{\text{fg}}(t_1, t_2, \dots, t_n)$	$\prod_{i=1}^{n} p_{\text{fg}}(t_i \vert T \setminus t_i, C)$	$\prod_{i=1}^{n} \frac{1}{2} \sum_{j\in\{1,-1\}} p_{\text{fg}}(t_i \vert T \setminus \{t_i, t_{i+j}\}, C)$
background	$p_{\text{bg}}(t_1, t_2, \dots, t_n)$	$\prod_{i=1}^{n} p_{\text{bg}}(t_i \vert T \setminus t_i, C)$	$\prod_{i=1}^{n} \frac{1}{2} \sum_{j\in\{1,-1\}} p_{\text{bg}}(t_i \vert T \setminus \{t_i, t_{i+j}\}, C)$

Table 5.1: Comparison of the probabilities from Tomokiyo and Hurst [2003] to the probabilities for the ideal and the simplified adaptation to am MLM. T are the tokens of a phrase to be scored, and C is the context surrounding the phrase. $t_1 \dots t_n$ are the tokens of a potential phrase with length n. The conditions of the conditional probabilities of the adapted formulations denote all the tokens of the sentence that are not masked. The formulation for the practical adaptation is not adequate, as it does not reflect tokens in boundary positions. For said tokens with only left or right context, only the existing context is used, without averaging.

Ideally, I wish to use the following reformulation as in column 'Ideal Adaptation' in Table 5.13. Instead of using the empirical probability of a phrase $p(t_1, t_2, ..., t_n)$, the probabilities of the single tokens of the phrase are multiplied given all other tokens of the phrase are visible $p(t_1|t_2, ..., t_n, C)$. And instead of using the phrase-independent token probabilities, the probabilities of the tokens are multiplied, but now with the other tokens of the phrase being masked: $p(t_1|C) \times \cdots \times p(t_n|C)$. For the case of keyness, the phrase's probability is compiled the same way for the background and the foreground model, with the phrase-dependent probability. To mitigate the influence of the tokens outside of the phrase (C), which are always unmasked and used by the model to infer the masks, the probabilities of all occurrences of the phrase are being averaged.

However, this presents a problem in terms of practicality. Let us assume the following sentence $\boxed{\text{A B C D E F G}}$ with length 7, presented in Table 5.2. The maximum phrase length is set to 6 to limit the computational cost, just as in the real experiment. Six sliding windows are applied to the phrase, so that all probabilities necessary in Table 5.13 can be extracted. The probabilities necessary for the phraseness of the phrase $\boxed{\text{B C D}}$ come from the masked sentences $\boxed{\text{A } \circ \text{ C D E F G}}$ $\boxed{\text{A B } \circ \text{ D E F G}}$ $\boxed{\text{A B C } \circ \text{ E F G}}$ and $\boxed{\text{A } \circ \circ \circ \text{ D E F G}}$. When setting a maximum phrase length, the computational complexity is linear. However, the memory complexity becomes problematic when more than one occurence of a phrase needs to be considered.

Full sentence	A B C D E F G		
window$_{l=1}$	○ B C D E F G	A ○ C D E F G	...
window$_{l=2}$	○ ○ C D E F G	A ○ ○ D E F G	...
...			
window$_{l=6}$	○ ○ ○ ○ ○ ○ G	A ○ ○ ○ ○ ○ ○	
Inferred sentences	$27 = \sum_{i=1}^{6} l_{\text{sentence}} + 1 - i$		
Extracted probabilities	$77 = \sum_{i=1}^{6} (l_{\text{sentence}} + 1 - i) * i$		

Table 5.2: Example of masking procedure

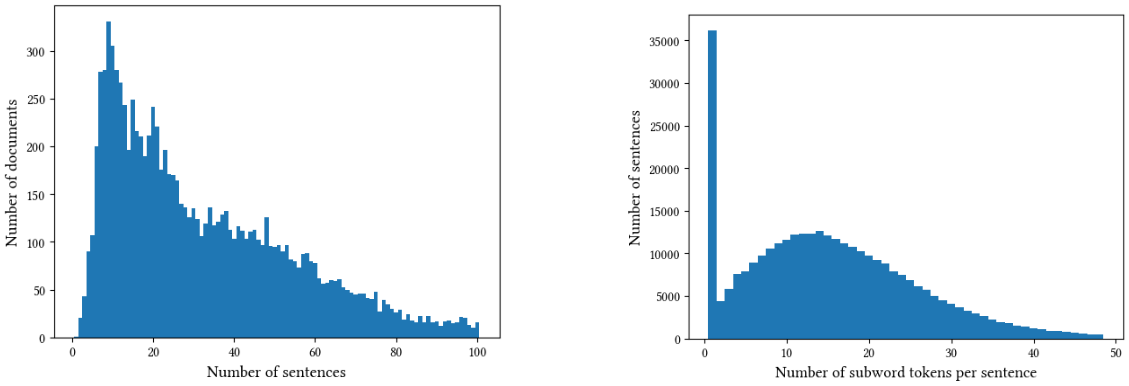
Please consider the masked phrase A B C ○ E F G. The probability for D needs to be retrieved for the calculation of all potential phrases inside this phrase containing D, e.g. C D D E F and B C D E, but not for any other phrases such as D X or Y D E F. Therefore, a rather unwieldy index structure needs to be maintained to access the probabilities.

This lead me to only evaluate the approach above for a sample of a 1000 documents of the foreground corpus. In another formulation that is easier on the hardware, I only keep track of the immediate left or right context of a mask. This means that the probabilities of tokens that appear in the middle of a phrase will not be accurately retrieved, there is always the chance that more probabilities are factored in. To arrive at aggregated probabilities one of those middle tokens, I average the probabilities of the token given the left context and the probability given the right context (‘Simplified Adaptation’, rightmost column in Table 5.13).

5.1.1.3 Processing Pipeline for Inference and Keyphrase Extraction

As introduced in Chapter 4, the articles of the January 2023 were filtered based on whether they were duplicates or not. Additionally, to reduce the inference time, I filtered out articles that contained either more than 100 sentences or a sentence with more than 50 subword tokens. From the remaining articles, I sampled 10,000 on which I inferred masks according to the scheme above. The probabilities were then entered into a datastructure that allowed for the quick access of all probabilities relevant to a phrase.

Considering the distributions in Figure 5.1, these filtering criteria appear to only have cut the tapering tail of the distributions, keeping the biggest part of the distributions intact.



(a) Distribution of document lengths for the foreground corpus sample

(b) Distribution of the sentence length in subword tokens for the foreground corpus sample

Figure 5.1: Distributions of the foreground corpus sample

5.1.1.4 A Word on Subwords

The adapted measures introduced above measure how strongly tokens inform each other. If the probability of a token gets lower, once its context is masked, we can assume that the masked tokens informed the model. If the probability drops severely, then there must be a strong association. On the other hand, if the probability increases, the context must be confusing the model, i.e. the model is not expecting the token to appear in this context.

It can be hypothesized that words that are part of idiomatic phrases would experience such a probability drop without their context. However, in the context of neural language models, the models don't infer on the words as such but on subword tokens. I expected the highest dependencies to be among subword tokens of the same word, as there are not only syntactic and semantic, but even morphological connections. This hypothesis was backed by preliminary experiments, where phraseness rankings were dominated by single words that were split into subword tokens. Kauf and Ivanova [2023] arrive at a similar finding, although in the context of pseudo-perplexity. There, subword tokens help a model to arrive at low pseudo-perplexities, if the method does not account for linguistic tokens.

For these reasons, words that are split into subwords are not considered for KE. Because the training of the tokenizer already considers the distribution of the data, we can expect that most words that are important to keyphrase extraction remain intact after subword tokenization. This is even more the case for this work, as the vocabulary of the tokenizer and subsequently the model was created on the same Swissdox data on which the background model was trained. However, it is still less then ideal, because KE is especially interested in what sets the foreground corpus apart. If the foreground corpus introduces new topics and accompanying lexical material which does get split into subwords, the approach is blind towards that. This is especially true for international news or sports reporting where many named entities do not necessarily follow the character distributions expected by a tokenizer trained on Swiss languages. It introduces a bias towards germanic and romance words, which would be highly problematic when deploying the approach in practice and outside of exploratory, methodological research.

Another thing to be wary of is the tokenizing procedure itself. The *SwissBERT* tokenizer is a retrained *XLM-R* tokenizer, which is in turn a *SentencePiece* tokenizer with special settings. What sets this tokenizer apart, is how it treats whitespaces. Whereas the training of other tokenizers employs pretokenization, *XLM-R* trains directly on the raw sequence [Conneau et al., 2020]. This can lead to whitespace tokens, like in this example: “Der Verkehrsminister steht unter öffentlichem Druck” gets tokenized to $\boxed{\text{\#Der}}$ $\boxed{}$ $\boxed{\text{Verkehrsminister}}$ $\boxed{\text{\#steht}}$ $\boxed{\text{\#unter}}$ $\boxed{}$ $\boxed{\text{öffentlichem}}$ $\boxed{\text{\#Druck}}$. For the words “Verkehrsminister” and “öffentlichem”, the tokenizer does not indicate the word boundary as part of the token, but as a secondary token. For consistency, I filtered out phrases that contained whitespace tokens, but the high dependence on them is worthy of investigation.

5.1.2 Results

This section presents the keyphrases extracted by the ideal and simplified adaptation. Additionally, the results of a direct replication of the original approach by Tomokiyo and Hurst [2003] are also presented for comparison[1], as well as keyphrases extracted by the LLR of corpus linguistics.

In Table 5.3 (ideal), Table 5.6 (simplified) and Table 5.9 (original), bigrams are ranked by their combined phraseness and informativeness score, which are also listed on their own. Trigrams are shown in Table 5.4 (ideal), Table 5.7 (simplified) and Table 5.10 (original). Table 5.5 (ideal), Table 5.8 (simplified) and Table 5.11 (original) show the same for 6-grams, the maximum phrase length. In the rightmost columns, the absolute frequency in the foreground corpus and in the sampled background corpus are listed. I choose the absolute over the relative frequency so one can directly read how many occurences were used to aggregate the probabilities (for the adapted cases).

[1] Tomokiyo and Hurst [2003] use Katz’ smoothing for unseen tokens in the background, where as I resorted to the simpler “Add δ” smoothing Chen and Goodman [1999]

	Score	I	P	$\#_{fg}$	$\#_{bg}$
Heimweg antreten	32.37	0.03	32.34	1	5
und entschied	31.39	-0.02	31.41	5	85
Robert Lewandowski	31.00	0.04	30.95	3	40
à peu	30.40	0.04	30.35	1	8
Diego Maradona	29.35	0.44	28.91	3	44
Breel Embolo	29.07	6.59	22.48	4	84
Glacier 3000	28.90	0.08	28.82	1	15
liebe und	28.54	0.08	28.45	1	17
Feuer fing	27.64	0.04	27.60	2	12
Burkina Faso	27.18	0.00	27.18	1	140
de facto	27.11	0.08	27.03	19	161
auslaufen lassen	26.78	0.07	26.71	3	17
Google Earth	26.27	0.21	26.06	1	17
gehalten »	25.94	0.62	25.32	6	29
to date	25.57	0.01	25.56	1	11

Table 5.3: Top 15 2-grams ranked by the ideal adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
up to date	59.21	0.03	59.19	1	9
neuem Leben erwacht	57.47	-0.11	57.58	1	9
Herrn und Frau	54.80	-0.03	54.82	5	25
Gott sei Dank	52.29	0.00	52.29	2	86
betrug letztes Jahr	50.45	0.09	50.36	2	10
peu à peu	48.50	0.04	48.46	1	8
in neuer Zusammensetzung	48.39	-0.03	48.43	3	19
ans Herz gewachsen	47.93	0.04	47.89	2	52
und interpretiert diese	47.18	-0.10	47.28	8	8
Museum Franz Gertsch	46.29	0.15	46.14	2	24
Grand Old Party	44.50	0.02	44.48	4	11
Amt und Würden	43.62	0.00	43.61	1	16
Schweizerin Belinda Bencic	43.18	0.01	43.17	3	0
weist vor Gericht	42.63	0.04	42.60	2	0
waschen à la	42.62	0.05	42.56	1	0

Table 5.4: Top 15 3-grams ranked by the ideal adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
nach wie vor davon ausgehen ,	79.56	0.04	79.53	2	1
Schülerinnen und Schüler unter einem Dach	75.35	0.08	75.28	1	0
wegen versuchten Mordes vor dem Bezirksgericht	71.78	0.08	71.70	1	0
Resultate und Kalender Resultate und Kalender	70.99	0.17	70.82	1	0
zur Gleichstellung von Mann und Frau	69.33	0.02	69.31	1	2
unter Berufung auf das Weisse Haus	68.61	0.02	68.60	2	0
« von A bis Z »	68.03	0.01	68.02	1	0
sorgt dafür , dass Granit Xhaka	67.95	0.16	67.78	1	0
habe sich in die Länge gezogen	67.68	0.01	67.67	1	0
scheint es sich auf den ersten	67.25	0.00	67.25	1	0
Tatsache , dass sich Marco Odermatt	66.36	0.70	65.66	2	0
, wo im Museum Franz Gertsch	64.63	0.15	64.49	2	0
was dazu führen kann , dass	62.81	0.02	62.79	1	2
als Tropfen auf den heissen Stein	62.60	0.02	62.58	2	2
ein Verstoss gegen das humanitäre Völkerrecht	62.29	0.03	62.26	1	0

Table 5.5: Top 15 6-grams ranked by the ideal adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
Sean Penn	37.27	0.39	36.88	1	26
Heimweg antreten	32.37	0.03	32.34	1	5
wenigsten am	30.67	0.02	30.65	1	2
gegründeten und	30.64	0.06	30.57	1	6
à peu	30.40	0.04	30.35	1	8
Gestalt annehmen	30.24	0.02	30.22	1	8
berichtete und	29.57	0.07	29.50	3	4
Bruno Ganz	29.56	0.03	29.53	2	32
Enfant terrible	29.53	0.19	29.33	1	36
Glacier 3000	28.90	0.08	28.82	1	15
Feuer fing	27.64	0.04	27.60	2	12
Mohamed Salah	27.48	-0.05	27.53	1	33
Burkina Faso	27.18	0.00	27.18	1	140
zu Mal	26.95	0.00	26.94	1	14
Google Earth	26.27	0.21	26.06	1	17

Table 5.6: Top 15 2-grams ranked by the simplified adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
Sieg um Sieg	27.13	-0.16	27.29	1	3
mit Füssen tritt	26.24	0.01	26.23	2	10
» Gestalt annehmen	25.81	0.02	25.79	1	0
vom hohen Ross	25.73	0.03	25.70	1	10
de suite gewinnen	25.62	0.50	25.12	1	0
lässt grüssen –	25.17	0.01	25.15	2	4
peu à peu	24.92	0.04	24.87	1	8
starkem Kontrast dazu	24.55	0.01	24.54	1	1
de suite steht	24.19	0.62	23.57	1	0
up to date	22.83	0.03	22.80	1	9
In Anlehnung daran	22.79	-0.01	22.80	1	0
wenig Zurückhaltung auferlegt	22.38	0.13	22.25	2	0
englische Fussballverband FA	22.33	0.24	22.10	1	1
mit Füssen getreten	21.63	0.07	21.56	4	29
à peu .	21.56	0.06	21.50	1	0

Table 5.7: Top 15 3-grams ranked by the simplified adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
Ryan Reynolds , Katie Holmes und	16.98	0.05	16.93	1	0
, Ryan Reynolds , Katie Holmes	16.60	0.04	16.56	1	0
Gotthard , Lötschberg und Monte Ceneri	14.09	0.33	13.75	2	0
Top Gun : Maverick » ab	13.99	0.71	13.28	1	0
gegen die Dallas Stars Moral bewiesen	13.98	0.13	13.85	1	0
, der es mit Füssen tritt	12.93	0.02	12.90	1	0
, Margot Robbie , Zlatan Ibrahimovic	12.33	0.17	12.15	1	0
Maus Frères in guten Händen .	11.87	0.04	11.83	2	0
Pink Floyd , Justin Bieber und	11.77	0.04	11.73	1	0
vor Ort , die Eindrücke vermitteln	11.60	2.56	9.04	50	0
de suite steht , kündigte an	10.95	0.33	10.62	1	0
bei Maus Frères in guten Händen	10.83	0.11	10.72	2	0
Pierre Maudet , Mauro Poggia ,	10.48	0.21	10.26	1	0
« Top Gun : Maverick »	10.21	0.80	9.40	1	1
auch Mujinga Kambundji ins neue Jahr	10.14	0.64	9.50	2	0

Table 5.8: Top 15 6-grams ranked by the simplified adaptation

	Score	I	P	$\#_{fg}$	$\#_{bg}$
, dass	0.01791	0.00034	0.01757	14294	166264
in der	0.01151	0.00017	0.01134	12753	153033
» ,	0.00881	0.00023	0.00858	10921	127940
, sagt	0.00674	0.00064	0.00610	5289	45627
. »	0.00664	0.00086	0.00578	9335	88126
, die	0.00620	-0.00049	0.00669	13496	188346
in den	0.00558	0.00001	0.00558	5882	73281
, wie	0.00524	0.00030	0.00494	5187	54292
: «	0.00514	0.00024	0.00489	4355	45983
für die	0.00498	-0.00010	0.00509	5737	75613
Bild :	0.00477	0.00089	0.00388	2105	7276
mit dem	0.00406	-0.00008	0.00415	3385	45652
« Ich	0.00375	0.00022	0.00352	2224	20432
« Wir	0.00351	0.00013	0.00338	2064	21303
Millionen Franken	0.00328	0.00006	0.00322	1366	14968

Table 5.9: Top 15 2-grams ranked by the original approach (without the noun phrase requirement)

	Score	I	P	$\#_{fg}$	$\#_{bg}$
» , sagt	0.01215	0.00038	0.01177	3809	34248
, dass die	0.00526	-0.00003	0.00529	2431	31482
in der Schweiz	0.00460	0.00011	0.00450	1539	15176
Exklusiv für Abonnenten	0.00291	0.00038	0.00253	429	300
» , sagte	0.00273	-0.00002	0.00275	941	12579
, dass es	0.00266	0.00007	0.00258	1051	10368
täglichen Update bleibst	0.00262	0.00036	0.00226	290	64
» , so	0.00257	0.00013	0.00244	982	7966
deine Lieblingsthemen informiert	0.00254	0.00036	0.00218	290	64
, heisst es	0.00251	0.00014	0.00237	732	4909
Januar 2023)	0.00247	0.00159	0.00088	202	0
Update bleibst du	0.00245	0.00036	0.00209	290	64
, dass sie	0.00239	0.00004	0.00235	917	9856
, dass er	0.00226	0.00004	0.00222	869	9173
knapp täglich direkt	0.00224	0.00036	0.00188	290	64

Table 5.10: Top 15 3-grams ranked by the original approach (without the noun phrase requirement)

	Score	I	P	$\#_{fg}$	$\#_{bg}$
Update bleibst du über deine Lieblingsthemen	0.00692	0.00030	0.00662	290	64
täglichen Update bleibst du über deine	0.00689	0.00030	0.00660	290	64
bleibst du über deine Lieblingsthemen informiert	0.00684	0.00030	0.00655	290	64
Mit dem täglichen Update bleibst du	0.00643	0.00030	0.00613	290	64
knapp täglich direkt in dein Postfach	0.00635	0.00030	0.00605	290	64
deine Lieblingsthemen informiert und verpasst keine	0.00629	0.00030	0.00599	290	64
dem täglichen Update bleibst du über	0.00629	0.00030	0.00599	290	64
Lieblingsthemen informiert und verpasst keine News	0.00618	0.00030	0.00589	290	64
über deine Lieblingsthemen informiert und verpasst	0.00616	0.00030	0.00587	290	64
Wichtigste kurz und knapp täglich direkt	0.00614	0.00030	0.00584	290	64
du über deine Lieblingsthemen informiert und	0.00599	0.00030	0.00569	290	64
keine News über das aktuelle Weltgeschehen	0.00586	0.00030	0.00556	290	64
verpasst keine News über das aktuelle	0.00579	0.00030	0.00549	290	64
News über das aktuelle Weltgeschehen mehr	0.00574	0.00030	0.00545	290	64
Erhalte das Wichtigste kurz und knapp	0.00571	0.00030	0.00541	290	64

Table 5.11: Top 15 6-grams ranked by the original approach (without the noun phrase requirement)

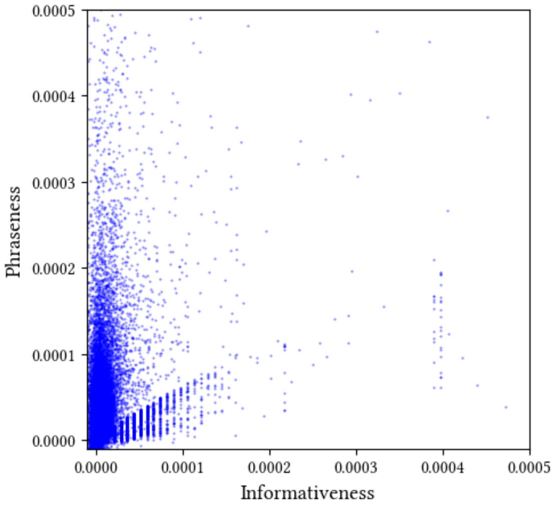
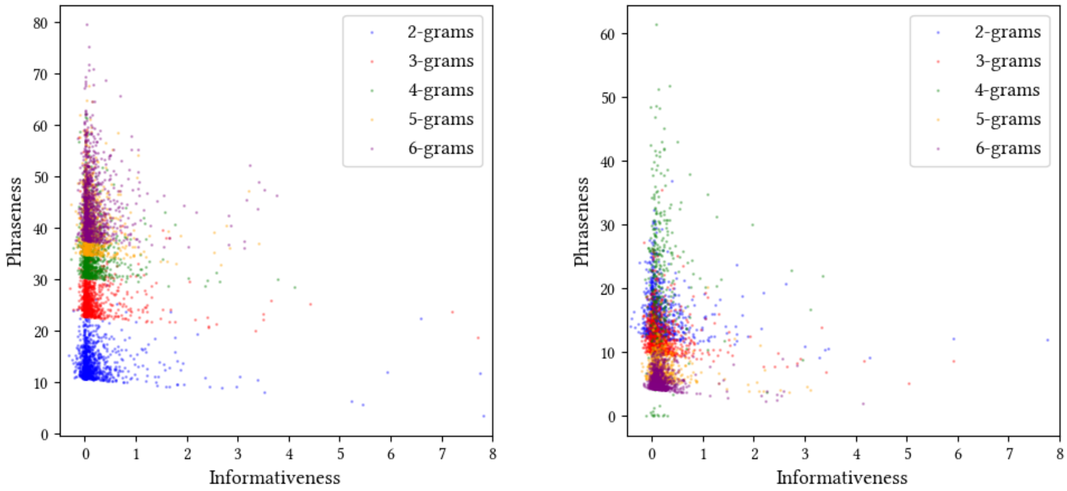


Figure 5.2: Informativeness and phraseness of all bigrams according to the original approach. Outliers are cropped.

The first thing to notice is that the keyphrases as such do not give a good indication of the thematic content of the foreground corpus, in neither of the tables. In the bigram and trigram tables however, one notices that some idiomatic phrases have very high ranks. Looking at the how the total score comes to be, these two points become clear. The phraseness is on a totally different scale than the informativeness, and dominates the score. The top 15 trigrams in all three methods even contain negative informativeness scores.

The original approach without any syntactic filter picks up lots of punctuation and stop words for bigrams (Table 5.9). Once again, the phraseness totally trumps informativeness, as can also be seen in Figure 5.2. The rising vertical lines in Figure 5.2 are explained by low-frequency events. Because many low-frequency bigrams appear with the same frequency, they subsequently get the same informativeness. A high informativeness indicates a larger relative frequency of the phrase in the foreground corpus, which translates to a higher phrase-dependent probability and subsequently a larger potential difference between phrase-dependent and phrase-independent probabilitiy. Therefore the horizontal bars appear to be rising with increasing informativeness.

Returning to the bigrams of the original approach in Table 5.9, and only considering it in terms of phraseness, the table can be explained. Token sequences as ". »" are expected to be highly associated, as all declarative sentences in direct speech would be terminated this way, a pattern very prevalent in news reports where the discourse of public figures is featured.



(a) Ideal Adaptation

(b) Simplified Adaptation

Figure 5.3: Informativeness and phraseness of the top 1000 n-grams

Moving up to tri- (Table 5.10) and 6-grams (Table 5.11), the original approach begins to suffer from its basis in frequency and duplicate sentences. Already present in the trigrams, the 6-grams feature 15 different slices of the same sentence, indicated by the equal number of appearances in the foreground and background corpus. Because some articles are always accompanied by the same banner, inflating the frequency of this specific word order, the original approach picks it up as highly associated.

The ideal adaptation to MLM, featured in Table 5.3 (bigrams), Table 5.4 (trigrams) and Table 5.5 (6-grams), shows the same behaviour between phraseness and informativeness as the original approach, but even more pronounced. The phraseness is several orders of magnitude stronger than the informativeness, and when increasing the phrase size, the differences become stronger (Figure 5.3a).

Interestingly, the opposite tendency is displayed by the simplified approach, where the phraseness trends lower and lower with increasing phrase size (Figure 5.3b). Remember that for the ideal approach, the phrase-dependent probability is the product of the single token's probability given that the whole rest of the sentence is unmasked, and therefore has a rather high probability. This is especially true after the model has been finetuned on exactly the sentences it is now inferring on. On the other hand, the phrase-independent probabilities come from tokens with their close context being masked, and the longer the phrase gets, the more potentially low probabilities get factored in, further decreasing the phrase-independant probability. This leads to a higher and higher pointwise Kullback-Leibler distance for phraseness. The same thing is not true for the simplified approach, where the amount of masked tokens remains the same, regardless of phrase size.

The bi- (Table 5.3) and trigram (Table 5.4) tables of the ideal adaptation show promising results in terms of phraseness. Alongside word combinations that do not appear to be particularly associated from an arm chair linguist's view, the approach picked up named entities such as "Robert Lewandowski", "Glacier 3000" (a ski resort), "Museum Franz Gertsch" and "Google Earth". But idiomatic phrases are shown as well, such as "Heimweg antreten", "Feuer fing", "ans Herz gewachsen" or "Gott sei Dank". Especially interesting are idioms that are borrowed from other languages like "peu à peu" (little by little), "de facto" and "up to date". Other trigrams that appear quite often, for both the ideal (Table 5.4) and simplified approach (Table 5.7), are so called hendiadys. Hendiadys is a term from rhetorics, describing a stylistic device, by which semantically similar or related concepts are conjoined by a coordinating conjunction. We find such hendiadys in phrases such as "Herrn und Frau" and "Amt und Würden", and not presented in the tables but still in the top 50 "Nadel und Faden" and "Höhen und Tiefen". The 6-grams don't really present idiomatic phrases as such, but often contain one or more idiomatic or closely associated words in subphrases: "$\big($nach wie vor$\big)$ $\big($davon ausgehen$\big)$," or "$\big($Schülerinnen und Schüler$\big)$ $\big($unter einem Dach$\big)$". Even for bi- and trigrams there are phrases that contain each other, exemplary "à peu" and "peu à peu". It is debatable to what extent "à peu" forms a phraseological unit by itself, and an idea to remedy this containment will be presented in the second experiment.

Still, for 6-grams (Table 5.5) , there are even idiomatic gems like "« von A bis Z »", "habe sich in die Länge gezogen" or "als Tropfen auf den heissen Stein". For phrases such as "Feuer fing" or "habe sich in die Länge gezogen" one needs to consider that these are actualizations of abstract patterns, that are not bound to a specific inflection of the verbs. Gries [2008] referred to this as the "flexibility" of phraseological units. However, considering the spectrum of substantive and formal idioms [Fillmore et al., 1988], these phrases are still on the side of substantiveness.

For informativeness, there is very little to work off, in all of the tables, already evident from the disproportionate relation to phraseness. The only phrase I can trace to the January of 2023 is "Museum Franz Gertsch", due to the death of its namesake on the 6th of January.

The tables for the simplified adaptation are similar to the ideal adaptation in lower phrase sizes, and get more and more different the longer the phrases become. This observation can be explained by the formulae being equal for the bigram case, and growing more and more dissimilar the longer the phrases become. Please remember that the differences in the lower tiers are also due to the different foreground corpus sizes (1,000 documents for the ideal approach, 10,000 documents for the simplified). It is also true for the 6-grams (Table 5.8) that they mostly are not phrasematic units themselves, but consist of several

Keywords according to Unigram Informativeness				Keywords according to Log-Likelihood-Ratio	
token	score	p_{fg}	p_{bg}	token	score
Lie	5.104839	0.731204	0.005787	2022	15637.920012
Pistorius	3.365052	0.393394	0.001047	2023	10940.560510
Lade	2.949329	0.728800	0.044095	-	10098.009223
Bilan	2.633663	0.740884	0.063047	Ukraine	4582.595301
verpassen	2.072911	0.602884	0.055615	2021	4253.127379
Ouest	2.022264	0.485291	0.027013	Januar	4081.136308
Eindrücke	1.826625	0.711012	0.119816	·	3611.157662
Affaire	1.635830	0.358307	0.015133	;	2925.598443
Esteban	1.557084	0.679925	0.139020	Berset	2757.909844
cycle	1.493981	0.623240	0.118318	News	2374.526124
Yvan	1.426630	0.608100	0.119604	2024	2255.605992
Elisa	1.313221	0.666378	0.170015	,	2249.572782
Cie	1.289857	0.860409	0.304386	Erhalte	2035.916999
zahlreich	1.283564	0.775632	0.246316	000	1903.748371
ONS	1.242676	0.957332	0.389321	Exklusiv	1795.049434
Amy	1.237957	0.594959	0.140644	Bild	1740.487208
hinterfragen	1.200385	0.476075	0.082919	bleibst	1667.100031
09:00	1.122314	0.116482	0.000146	Lieblingsthemen	1649.212761
2023	1.094497	0.379440	0.051383	Abonnenten	1601.888716
Augenzeugen	1.001479	0.529969	0.143020	–	1601.188123
Astronaut	0.985616	0.658413	0.233275	Weltgeschehen	1571.268516
DU	0.980886	0.437411	0.092435	du	1520.568213
logg	0.977633	0.634374	0.217986	.	1483.289115
widersprüchlich	0.954496	0.838274	0.380732	Lützerath	1480.443471
Bafu	0.862501	0.262544	0.026932	sagt	1415.626884
Società	0.858664	0.881862	0.449045	Dezember	1387.665921
VC	0.837612	0.616830	0.240651	Lauener	1379.310286
prüft	0.805055	0.492431	0.158564	Ski	1340.311991
Pietro	0.797201	0.257312	0.030048	Update	1328.785229
Valle	0.787204	0.886077	0.478666	Postfach	1285.955968

Table 5.12: Comparison of keywords according to the unigram informativeness and the log-likelihood ratio

highly associated words, mainly two-token named entities. This can be explained by the simplification, because the phrase-independent probabilities stem from bigram masks.

5.1.2.1 Informativeness on its own

To still get an insight into how informativeness behaves, let us consider how phrases are ranked only according to informativeness. The easiest way to do this is to consider unigrams. The method averages over all probabilites assigned to a token given full contexts, and then we compare the probabilities of the foreground model to those of the background model with the pD_{KL}. To have a reference, the key unigrams identified by the log-likelihood ratio (LLR) are displayed as well. Results are shown in Table 5.12

The keywords identified by the LLR appear to be a lot more indicative of the foreground corpus. Seasonal and temporal indicators exist as is expected for a foreground corpus temporally situated in the January of 2023, when the background corpus spans the past 30 years: “2021” - “2024”, “Januar”, “Dezember” and “Ski” all fall into this category. There are also a few words that appear to be (and after corpus inspection, mainly are) parts of article banners that are repeated among several articles, such as "News", "Exklusiv" or "Abonennten". Again, this can be attributed to the fact that the banners on news outlets did not change much during the month of January, but changed enough during the background corpus' timespan to be distinctive for the foreground corpus. Even topically, "Ukraine", "Lützerath", "Berset" and "Lauener" all point to topics that were central to public discourse in the January of 2023. "Ukraine" due to the ongoing war of Russia against Ukraine, "Lützerath" due to the occupation and subsequent eviction of the German village of the same name by climate activists, and "Berset" and "Lauener" due to an affair around the indiscretions between the department of health and the media.

On the other hand, the keywords identified by the unigram informativeness appear hermetic. At least two words are easily interpretable, evidentially "2023" and the second ranked "Pistorius". Boris Pistorius was appointed as the German minister for defence in the January of 2023, which was accompanied by broad media coverage at least partly due to the interest in the development of military relations between Germany and Ukraine. The top ranked "Lie" is part of the name of the norwegian ski athlete Atle Lie McGrath. "Lade" appears mostly in the imperative in "Lade hier deine Bilder hoch." a pattern that only makes sense for a news outlets that embraced web 2.0 with its paradigm of user-generated content, something that would not hold true for the most part of the background corpus' timespan. "Bilan" is the last name of photographer Clemens Bilan, who appeared twice as a source. A product of duplicate sentences is "verpassen", which occured 290 in the banner phrase "Keine News mehr verpassen". The same is true for "prüft", which appears 50 times in the phrase "Wie prüft SRF die Quellen in der Kriegsberichterstattung?". The background model proposes "beurteilt" instead. As would be expected, these duplicated occurences heavily bias the model learn the phrases itself, and even Tomokiyo and Hurst [2003] also note this phenomenon in their paper for the phrase "message news".

In general, when one considers the key tokens of the unigram informativeness, they appear to be unusual and rare. Tokens that clearly stem from languages other than German and abbreviations as "ONS", "DU", "Bafu" and "VC" are interspersed with tokens that appear to be common as "verpassen" and "prüft", but that proved to be part of duplicated sentences. The same phenomenon continues with n-grams of higher order, although not printed here. The problem of duplicate sentences can be remedied quite easily during preprocessing, but the other effect is inherent to the approach. The informativeness, when calculated with pD_{KL}, seems to favor tokens and combinations that are very unlikely in the background distribution. This then especially picks up phrases that contain tokens that are not part of standard German language. In the terms of the different aspects of keyness, the pD_{KL} for informativeness favors tokens that have a high distinctiveness in terms of language, while KE is interested in distinctiveness in terms of discourse.

5.1.3 Discussion

To first thing to state is that the direct adaptation of Tomokiyo and Hurst [2003] did not yield sufficient results for KE. For one, the scales of informativeness and phraseness are disjointed by several orders of magnitude. This holds true even for the direct replication of the original approach.

Even if one considered rescaling informativeness and phraseness for the adaptation, the problem of informativeness favoring rare tokens would persist. For further experiments, I would therefore start evaluating different ways of calculating informativeness, maybe even resorting to the established log-likelihood ratio.

Another thing to note is that I did not show the application of the hard filters of Tomokiyo and Hurst [2003], meaning I did not only report noun phrases with a minimum count of 5. Preliminary experiments showed how this did not improve on the results. For the clarity of the methods and the progression of my argumentation, I reported the results without a filter (except for whitespace tokens).

Opposed to informativeness, the phraseness score provided results that I deem worthy of further exploration. Using the probabilities assigned by an MLM given differently masked contexts proved to yield results that at least partially coincide with my judgement of what makes a phrasematic unit. In the further experiments, I will therefore focus on phraseness.

5.2 Phraseness

The previous experiment proved to yield poor results in terms of topical keyphrases. Instead of tweaking the methods e.g. through the introduction of learnable parameters to level the playing field for informativeness and phraseness, I want to explore the dimension of phraseness. This line of attack is motivated twofold: First, the previous experiment showed how the extraction of probabilities from MLMs with different configurations of masks can lead to the discovery of idiomatically associated words. An approach that focuses solely on the task of phraseness in the context of MLMs, leaving behind the ideas of Tomokiyo and Hurst [2003], may be even more fruitful. Secondly, and more abstractly, I see this line of experiments as a venture into the field of interpretability. It can yield insights into how MLMs associate tokens with each other, and whether there are patterns that are worthy of further investigation or even consideration when deploying MLMs.

5.2.1 Methodology

Please consider the following sentence: “You’re stressed, you need to blow off some steam”. It ends with the idiom “blow off some steam”. Now, assume “blow” is masked: “You’re stressed, you need to ○ off some steam”. Knowing the idiom, we would be able to guess the word with quite a high chance of being right. Masking “off” or even “to” would not really diminish our chances, but once “steam” is masked, our chances of guessing “blow” are dropping. “catch some fresh air”, “learn how to rest”, “find your spirit animal” are all potential fillers for “You’re stressed, you need to ○ ○ ○ ○”.

I hypothesize that the phraseness of a phrase can be determined by extending the masking window and keeping track of the probabilities of the boundary tokens assigned by an MLM. Once the probability drops, this indicates a high association and thus a high phraseness. This approach is rationalistically advantageous to the phraseness calculations of the ideal adaptation from the previous experiment, because the probabilities being compared don’t stem from multiplying individual probabilities. It is always the probability of the leftmost (or rightmost) token being compared to its own probability, but once more context is masked. On the other hand, the approach appears to be biased towards these boundary tokens, not really accounting for the inner dependencies for the phrase. I experimented with different ways to include the inner live of the potential phrases, but the results were inconclusive.

Once more, I am using the pointwise Kullback-Leibler divergence. However, I am also drawing upon the normalized pointwise mutual intelligence as formulated by Vamvas and Sennrich [2023]. It is important to note that Vamvas and Sennrich [2023] do not use the entropy H as defined in the chapter on Information Theory, but the cross-entropy of the predicted probability with the empirical probability (the correct token has a probability of 1). The cross-entropy then breaks down to the negative logarithm of the predicted probability for the token, because all other possible tokens would be multiplied with 0. Furthermore, Vamvas and Sennrich [2023] don’t compare the probability of token t_1 once another token is (un-)masked, but the probability of token t_1 once another sentence is appended to the context.

Please consider how, in information theoretic terms, two different things are quantified by the (n)pmi and the pD_{KL}. For the pointwise Kullback-Leibler divergence, we measure how many excess bits are being used when we would be encoding $p(t_1|t_n, C)$ based on an ideal code for $p(t_1|C)$. If $p(t_1|t_n, C)$ is larger than $p(t_1|C)$ then the value is positive,

	Formula
pD_{KL}	
Unidirectional	$Score_{lefttoright}(T) = pD_{KL}(p(t_1 \mid t_n, C) \Vert p(t_1 \mid C))$
	$Score_{righttoleft}(T) = pD_{KL}(p(t_n \mid t_1, C) \Vert p(t_n \mid C))$
Bidirectional	$Score = \frac{1}{2}(Score_{lefttoright} + Score_{righttoleft})$
npmi	
Unidirectional	$Score_{lefttoright}(T) = npmi(t_1 \mid C; t_n, C) = H(t_1 \mid C) - H(t_1 \mid t_n, C))$
	$Score_{righttoleft}(T) = npmi(t_n \mid C; t_1, C) = H(t_n \mid C) - H(t_n \mid t_1, C))$
Both directions	$Score = \frac{1}{2}(Score_{lefttoright} + Score_{righttoleft})$
Inner Dependencies	$\frac{1}{n} \sum_{T_{sub} \in T} Score(T_{sub})$

Table 5.13: Formulae used for the assessment of phraseness and idiomaticity

as lower probabilities translate to a longer ideal code, and therefore more bits would be used.

On the other hand, the pointwise mutual information (pMI) measures the amount of information shared by two events, i.e. how much the uncertainty about one event is reduced when knowing the other event. However, the pMI is not as straightforward to interpret in the terms of a MLM. One of the defining features of the 'abstract' pMI is its symmetry for the two events. The uncertainty of an event decreases the same amount knowing the other one as vice versa. This doesn't strictly hold for MLMs, as it would for n-gram language models. This will show when contrasting the scores for a phrase when comparing the npmi left-to-right to the npmi right-to-left. In information theory, they should be equal, in practice, they are not. That is why even for the npmi, I evaluated uni-directional scores and averaging over both reading directions.

To compare the formulae behind npmi (Equation (5.2d)) and pD_{KL} (Equation (5.1)), let me shorten $p(t_1|t_n, C)$ to p_{cond} and $p(t_1|C)$ to p_{prior}. After transforming the npmi (Equation (5.2f)), it becomes clear that the differences between the two formulae come down to a scaling factor. The pD_{KL} scales according to the conditional probability, i.e. the probability of t_1 when t_n is visible. The npmi however scales according to the larger entropy, i.e. the lower probability, which, for associated tokens, will be p_{prior}.

$$pD_{KL}(T) = p_{cond} \times \log_2 \left(\frac{p_{cond}}{p_{prior}} \right) \tag{5.1}$$

$$
\begin{aligned}
H(p_{prior}) &= \log_2\left(\frac{1}{p_{prior}}\right) &\text{(5.2a)}\\
H(p_{cond}) &= \log_2\left(\frac{1}{p_{cond}}\right) &\text{(5.2b)}\\
normalization &= \frac{1}{max\left(H(p_{prior}), H(p_{cond})\right)} &\text{(5.2c)}\\
npmi(T) &= \left(H(p_{prior}) - H(p_{cond})\right) \times normalization &\text{(5.2d)}\\
&= \left(\log_2\left(\frac{1}{p_{prior}}\right) - \log_2\left(\frac{1}{p_{cond}}\right)\right) \times normalization &\text{(5.2e)}\\
&= \log_2\left(\frac{p_{cond}}{p_{prior}}\right) \times normalization &\text{(5.2f)}
\end{aligned}
$$

Where pD_{KL} penalizes phrases where the conditional probability is not as high, the npmi (for phrases with positive phraseness, i.e. where the $p_{cond} > p_{prior}$) penalizes phrases where the prior probability is very small. This difference will prove to lead to more robust results for the npmi scoring mechanism.

Additionally, I noticed how the measures introduced above are quite sensitive towards the whitespace token of the *XLM-R* and subsequently *SwissBERT* tokenizer, even more so than before. This shows especially for bigrams, where the rankings are dominated by the "_ X" pattern. In a technical sense, this can be explained, as the presence of the token "X" almost always indicates the need for a whitespace token.[2] The measured dependency does not rely on linguistic connections, but on the relations of a processing artefact. To arrive at interpretable tables, I therefore again filtered out phrases containing whitespace tokens Furthermore, I needed to filter a few bigram cases (less than 10) where the data pipeline rounded the probabilities of highly likely tokens $p(t_1|t_2, C)$ to 1. For npmi, this always lead to maximum scores of 1, due to the entropy of t_1 effectively decreasing to 0 once t_2 is known.

The following results stem from the probabilities assigned by the finetuned foreground model to the same 10,000 document sample as the previous experiments. The probabilities of the same phrases are aggregated as in the simplified approach, but now with keeping track of a single unmasked token to the right and left of the masking window.

[2] To investigate this, I searched the *SwissBERT* tokenizer's vocabulary for tokens that only appear without a word boundary indicator in front. These tokens include words as "Debakel", "Beschwerdeführer", "Arizona" or "Fiction". Whenever these words appear with a whitespace in front, and my approach masks this whitespace token and evaluates e.g. "Debakel" as context, it can predict the whitespace with a very high certainty. In settings where both tokens are masked, more probable words are preferred, making the prior probability of the whitespace token very small.

Left-to-right			Both directions		
	score	$\#_{fg}$		score	$\#_{fg}$
Innern damals	0.999998	1	Oskar Freysinger	0.999988	1
Aston Martin	0.999997	1	Hong Kong	0.999973	2
SLF gibt	0.999997	1	Marin Cilic	0.999966	2
Schnegg (60	0.999993	3	Sierra Leone	0.999957	2
Frères wolle	0.999993	1	Buenos Aires	0.999951	11
AS Roma	0.999992	3	Pfund Sterling	0.999926	2
Guardia di	0.999991	1	Karolina Pliskova	0.999925	1
Atalanta Bergamo	0.999989	2	Mona Lisa	0.999913	1
sich Davos	0.999989	1	Billie Jean	0.999906	1
zu verstossen	0.999988	2	Brunschwig Graf	0.999905	1
Evi Allemann	0.999988	1	Victoria Jungfrau	0.999901	1
suite gewinnen	0.999988	1	unsicher gemacht	0.999900	1
Olympique Marseille	0.999986	1	Sebastian Vettel	0.999900	1
Augen führe	0.999985	1	Martina Hingis	0.999882	1
) Rohöl	0.999985	1	kurz oder	0.999865	2
Leutenegger Oberholzer	0.999985	1	Jens Stoltenberg	0.999865	14
Kathrin Bertschy	0.999984	1	Ajax Amsterdam	0.999859	4
gerufen habe	0.999982	1	zu Mal	0.999859	1
Füssen getreten	0.999982	6	Blerim Dzemaili	0.999850	1
Oskar Freysinger	0.999981	1	Simona Halep	0.999849	1

Table 5.14: Top 20 2-grams with left-to-right npmi (left) and bidirectional npmi (right)

5.2.2 Results

I will be presenting the results of the approaches mentioned above in order of increasing complexity. This also mimics my development journey, where I tried to the remedy the shortcomings of earlier approaches by introducing new terms or concepts.

5.2.2.1 Comparison of Unidirectional and Bidirectional Scoring

Table 5.14 (bigrams), Table 5.15 (trigrams) and Table 5.16 (4-grams) show the top 20 of the rankings for the left-to-right unidirectional npmi and the bidirectional npmi. Looking at the bigram table, one notices how both approaches feature named entities: People, places and sports clubs, or a currency, as in “Pfund Sterling”. However, the bidirectional approach appears to be more consistent, with almost no apparent ‘false positives’ in terms of my judgement of association. Interesting, and not necessarily ideal, are the high ranks of subphrases even for the bidirectional approach. Please consider the phrase “kurz oder” on rank 15 for the bidirectional approach. It appears twice in the corpus, and twice in the known idiom “über kurz oder lang”. The full idiom appears in rank 299,990 of 1,478,796 ranked phrases of size four, ranking behind other slices of the phrase such as “führen über kurz oder”. Somehow the bidirectional npmi picks up a strong association somewhere in the phrase, but not for the whole phrase as such. However, there are still actualizations of idiomatic phrases of length two, such as “unsicher gemacht” or “zu Mal”.

Left-to-right			Both directions		
phrase	score	$\#_{fg}$	phrase	score	$\#_{fg}$
Stimmbürgerinnen und Stimmbürger	0.999991	21	als 1000 Zeichen	0.999924	1
Einzug zu halten	0.999991	1	Einzug zu halten	0.999908	1
fort und stiegen	0.999987	2	von Zeit zu	0.999847	1
gut und ordnete	0.999984	2	Termin zu Termin	0.999807	2
auf stärkere Beine	0.999980	1	" Blick "	0.999803	3
Hochburg der Terroristen	0.999979	1	Dorn im Auge	0.999779	21
Grenzen : «	0.999979	1	« Zimmer »	0.999742	1
vor 2 Stunden	0.999976	2	zwischen Madrid und	0.999738	1
erlag der Radfahrer	0.999967	1	« Winter »	0.999731	1
von Zeit zu	0.999964	1	« erheblich »	0.999727	1
, die nachts	0.999960	2	Lokal zu Lokal	0.999727	2
und interpretiert diese	0.999959	8	fing ich an	0.999704	1
Innern zu entnehmen	0.999957	1	als 1500 Zeichen	0.999691	6
als 1000 Zeichen	0.999955	1	für Chaos gesorgt	0.999655	1
von Minute zu	0.999954	2	Teppich zu kehren	0.999622	2
» in Kolumbien	0.999951	2	Gruppe zu Gruppe	0.999613	1
Veröffentlichung des Songs	0.999948	1	Fall zu Fall	0.999612	1
zu ihren Gunsten	0.999945	7	« fern »	0.999574	1
, Gilles Marchand	0.999943	1	Osama Bin Laden	0.999557	1
zu einer bis	0.999943	2	Katz und Maus	0.999555	1

Table 5.15: Top 20 3-grams with left-to-right npmi (left) and bidirectional npmi (right)

Left-to-right			Both directions		
phrase	score	$\#_{fg}$	phrase	score	$\#_{fg}$
gewandt , berichtet Stark	0.999990	1	um an ihn zu	0.999577	1
dar », sagt er	0.999988	1	äussert sich an dieser	0.999263	1
Alter von 91 Jahren	0.999987	1	zwischen dem 16. und	0.999262	2
Alter von 84 Jahren	0.999985	3	aus dem Urteil hervorgeht	0.999238	1
wie die SNB mitteilte	0.999980	1	geht es mehr darum	0.999190	1
, die Israel und	0.999975	1	Alter von 91 Jahren	0.999035	1
um einen Unfall oder	0.999974	10	Alter von 84 Jahren	0.999019	3
Alter von 85 Jahren	0.999964	1	wie die Lausanner bekannt	0.999016	1
Rede von 15 Jahren	0.999963	1	aus der Antwort hervor	0.998928	1
Alter von 68 Jahren	0.999959	1	kam , wie es	0.998920	1
Meillard und Daniel Yule	0.999958	1	aus der Weisung hervorgeht	0.998908	1
vor , im neuen	0.999957	1	aus einem Bericht hervor	0.998808	3
konfrontiert , er sei	0.999955	3	Alter von 85 Jahren	0.998726	1
, welche 15 Kilometer	0.999953	1	mit einer Strafe rechnen	0.998706	1
, die der Schweizer	0.999949	1	" Neue Klasse ",	0.998635	1
Alter von 89 Jahren	0.999949	1	Alter von 68 Jahren	0.998586	1
wie das Unternehmen mitteilte	0.999943	1	aus dem Baugesuch hervorgeht	0.998559	1
Alter von 42 Jahren	0.999937	1	– etwa Italien –	0.998543	1
her und gibt sich	0.999937	1	Alter von 75 Jahren	0.998459	3
es aus zur Kollision	0.999935	1	Wert von 40 Franken	0.998359	1

Table 5.16: Top 20 4-grams with left-to-right npmi (left) and bidirectional npmi (right)

Considering the trigram table (Table 5.15), other patterns start to emerge. For one, there are still subphrases as “von Zeit zu” of the full idiom “von Zeit zu Zeit”. Once again, the full idiom lies far behind in the ranking, this time even with a negative npmi score. Prominently featured in the bidirectional ranking are single words in quotes. This is very easily explained, as the presence of an opening quotation mark implies the presence of a closing one, and vice versa. The chance of an opening quote apparently increases dramatically when a closing quote is unmasked. Interestingly, actualizations of the same abstract phrase start appearing, for example with “als 1000 Zeichen” and “als 1500 Zeichen”. When looking at the corpus, a human judge would probably identify “mehr als X Zeichen” as an abstract phrase. This pattern appears in a formulaic fashion when a newspaper asks readers to write a letter to the editor. Another, not fully extended and differently actualized pattern is “von X zu X”, as in the already mentioned “von Zeit zu“, but also in “Termin zu Termin”, “von Minute zu”, “Fall zu Fall” and “Lokal zu Lokal”. Once again, there are also idioms identified without any further objections, such as “Einzug zu halten”, “zu ihren Gunsten”, “für Chaos gesorgt”, “Dorn im Auge” or the hendiadys “Katz und Maus”.

The 4-grams in Table 5.16 continue showing the same characteristics. Some patterns are actualized multiple times: “aus dem Urteil hervorgeht”, “aus der Antwort hervor”, “aus der Weisung hervorgeht”, “aus einem Bericht hervor” and more. Another pattern is “Alter von X Jahren”, a highly frequent pattern, mainly used to talk about the age of recently deceased people. In the foreground corpus it is almost always used with the preposition “im”, once with an adjective, “im zarten Alter”. Once again, the 5-grams of “im Alter von X Jahren” appear nowhere close to the top of the rankings. This might be due to the fact that “im Alter von X” also appears with other continuations than “Jahren”, and therefore the informative content of “Jahren” is not as big as it would be.

The characteristics shown so far keep up in the 5- and 6-gram rankings, although not shown here. For 5-grams, prominent patterns in the top 30 are “nicht nur X , sondern”, “weder X Y , noch” or “Höhe von X Y Franken“.

For completeness sake, see Table 5.17 for the comparison of trigrams ranked by the unidirectional and bidirectional pD_{KL}. Especially the bidirectional case is not dissimilar to the bidirectional npmi. Words in single quotes and several hendiadys appear as well. The unidirectional method also shows some of the same phrases as the bidirectional one, but also many phrases that do not immediately become recognizable as patterns.

Comparing npmi (Table 5.15) and pD_{KL} (Table 5.17), the npmi appears to produce results that are better interpretable in terms of known patterns or idioms. Recurring to my interpretation of the difference between npmi and pD_{KL} in the methodology chapter, the scaling undertaken by npmi, i.e. a penalty for a low prior probability (npmi) produces

Unidirectional			Bidirectional		
phrase	score	$\#_{fg}$	phrase	score	$\#_{fg}$
blickt der gebürtige	15.821901	2	" besondere "	12.900713	1
aus Kosovo stammende	14.246228	1	und zu unterstützen	11.541717	2
" besondere "	13.650791	1	Amt und Würden	11.324233	1
finden dieses Jahr	13.600592	1	in Barcelona verurteilte	11.148547	1
meldete vergangene Woche	13.453090	1	Katz und Maus	10.586514	1
Erdgas und kommt	13.081240	1	Spital zu Spital	10.438927	3
übt seit Oktober	13.010704	1	« siedeln »	10.245218	1
ruft die Organisation	12.826321	1	aus Kosovo stammende	10.238107	1
sage und schreibe	12.756336	3	« Milliarden »	10.225139	1
Händler und Marktteilnehmer	12.252583	8	Schulter an Schulter	10.195337	2
fällt ein Jahr	11.949694	2	Händler und Marktteilnehmer	10.142287	8
es gehe ihnen	11.896095	1	gilt , teilgenommen	10.005190	1
Opfer eines Angriffs	11.787164	1	zweiter und dritter	9.941329	3
in Barcelona verurteilte	11.422213	1	« Sponsor »	9.909557	1
machte am Montagabend	11.291347	1	« Löwen ».	9.887259	1
« Löwen ».	11.221874	1	weder Polizei noch	9.778055	1
Amt und Würden	11.190802	1	sage und schreibe	9.580067	3
geheiratet , haben	11.135930	1	« Point »,	9.490280	1
stellt zum Beispiel	11.112356	1	stattfindet , bekannt	9.363297	1
Marine Le Pen	11.100678	2	stattfindet , bekannt	9.363297	1

Table 5.17: Top 20 3-grams for unidirectional and bidirectional pD_{KL}

better results than a penalty for low conditional probabilities (pD_{KL}). This might be due to a smoothing effect that occurs with the scaling factor of npmi, because the influence of tokens with very low prior probabilities are dampened. This is visible in Figure 5.4, where pD_{KL} essentially only assigns high scores to combinations with extremely high conditional probabilities and very low prior probabilities. The npmi is much more relaxed in this regard, which appears to benefit the rankings according to my judgement. For the further experiments, I will therefore focus on npmi.

5.2.2.2 Integration of Subphrase Scoring

In the last chapter, at several places I noted how phrases that are part of a bigger idiom are ranked high in their respective table, but the full idiom is ranked far behind. The main idea to remedy this is based on integrating the scores of subphrases into the score of a phrase. This would not counter the high ranking of subphrases, but it might increase the rank of the whole phrases.

Regardless of the findings of the last experiments, it appears to be logical to try to include the internal dependencies of a phrase into its score. I experimented with different ways to include the phraseness of the subphrases, from different aggregation methods as well as treating the 'subphraseness' as a bonus or malus.

The first way to calculate a subphraseness score consists of calculating the previously introduced bidirectional npmi for all subphrases of a phrase and averaging over them.

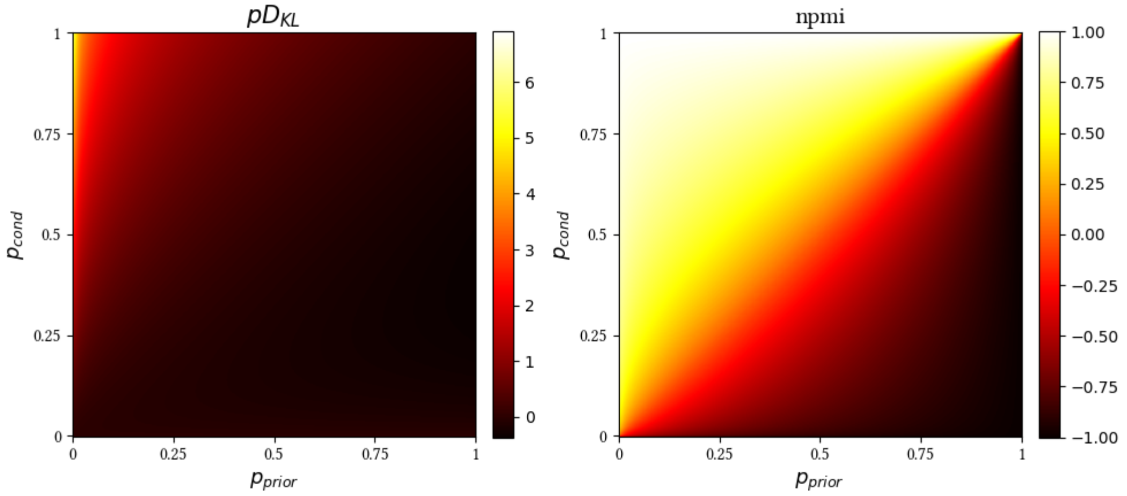


Figure 5.4: Comparison of pD_{KL} and npmi

Please note how the probabilities of the same subphrases that are not occuring in this specific phrase in the corpus are added into the aggregation as well. This means that when calculating the subphraseseness for "B C" when assessing the full phrase "A B C", probabilities for "B C" in sentences such as "X B C" will be part of the subphraseness. This subphraseness score is then simply added or subtracted (depending on treating it as a bonus or malus) to the bidirectional npmi of a phrase according to the last section. Because the subphraseness is recursively calculated, the effect of subphrases gets compounded.

Table 5.18 presents 4-grams when compiling the subphraseness by averaging over the subphraseness scores and either treating it is a bonus or malus, i.e. adding or subtracting it from the score of the boundary tokens of the full phrase. Looking at subphraseness as a bonus, and also comparing these results to Table 5.16 where subphraseness isn't included, one can see how there are no longer multiple phrases with different fillers. Named entities start reappearing, whereas the bidirectional npmi no longer ranked them as highly for phrases longer than 2 (with the exception of "Osama Bin Laden". Other patterns are verbal brackets with reflexive pronouns ("habe ich mich gefühlt" and "habe ich mich entschieden") or verbs with idiomatic associates. Once again, many of the listed phrases are subphrases of larger idioms, for example in "Strich durch die Rechnung" (with "machen"). "Kauf nehmen zu müssen" and "Angriff genommen werden können" are both part of a phraseological unit that is usually prefixed by an "in". Apparently, the dependence is even stronger among the verbal components.

Something that gets picked up by the subphraseness bonus, that might not by ideal, is visible in the phrase "auf künstliche Intelligenz setzt". The German language has verbs

Bonus for high subphraseness			Malus for high subphraseness		
phrase	score	$\#_{fg}$	phrase	score	$\#_{fg}$
New York Stock Exchange	2.193635	1	« machen können »	1.098344	2
habe ich mich gefühlt	2.067913	1	um Punkt zwölf Uhr	1.082386	1
habe ich mich entschieden	2.058119	3	schliess der Konzern nicht	1.078573	1
einem sehr frühen Stadium	2.058002	2	finnischen Meister in Tampere	1.076984	1
Strich durch die Rechnung	2.049010	13	unserem Beschluss strikt fest	1.074988	1
portugiesische Superstar Cristiano Ronaldo	2.046179	1	zwischen Oktober 2021 und	1.074500	1
die Vereinigten Arabischen Emirate	2.033379	2	über Richterswil gut informiert	1.073432	1
egal wie alt sie	2.029040	1	Peter Lüscher und Peter	1.071830	1
Art und Weise kennen	2.011835	1	zwischen der ukrainischen und	1.066805	1
« Im Hinblick auf	2.009694	1	als sonst schon .	1.062595	1
blieb beim Unfall unverletzt	2.009538	1	nimmt 6. Titel ins	1.058042	1
gesundheitlichen Gründen nicht mehr	2.006884	2	– zurzeit zumindest –	1.050835	1
Kauf nehmen zu müssen	1.990773	3	wurde erst 1966 gegründet	1.045959	1
Angriff genommen werden können	1.988629	1	und nicht überfordert werden	1.044103	1
auf künstliche Intelligenz setzt	1.982565	1	zwischen wichtigen Investitionen und	1.042823	1
Zenit St . Petersburg	1.982265	1	auf knapp 100'000 Franken	1.042626	1
« Harry Potter »	1.962029	2	, diese vorzunehmen .	1.042031	1
unter Beweis zu stellen	1.958625	2	um an ihn zu	1.041406	1
, Papst Franziskus ,	1.958192	1	dass sie lange Freude	1.040361	1
St . Galler Tagblatt	1.956815	6	für 3,4 Milliarden Dollar	1.038535	1

Table 5.18: Top 20 4-grams when using the subphraseness score (averaged) as a bonus and as a malus

that occur in so called 'festen Verbindungen', in fixed conjunctions, with prepositions determining the case of the verb's object, "setzen auf" is an example for that. These specific verb and preposition patterns are something we might expect to be returned from all the MLM phraseness scoring approaches. The subphrase "künstliche Intelligenz" has in itself a high phraseness, but not because it is part of an idiom with the surrounding verbal group, but because it is an established phrase by itself. The same could be the reason for "« Harry Potter »" or the inserted nominal phrase ", Papst Franziskus ,". "Papst Franziskus" is in itself a named entity, and given the context of the rest of the phrase, the two commas determine each other. Two highly associated token pairs appear together, and the subphraseness bonus lifts the combination to the top of the ranking.

The subphraseness as a malus gives interesting results, seen in the right side of Table 5.18. Please focus only on the first and last word of the identified phrases. Some of the phrases then exhibit strong connections, semantically, syntactically and lexically. A semantic connection is exhibited in the fourth rank with "**finnischen** Meister in **Tampere**", whereas Tampere is a city in Finland. Syntactical connections are displayed once again with quotation marks and dashes, but also with phrases as "**zwischen** Oktober 2021 **und**". Lexical connections such as "**um** Punkt zwölf **Uhr**", "**nimmt** 6. Titel **ins**", "**auf** knapp 100'000 **Dollar**" or "**für** 3,4 Milliarden **Franken**" present associated words that form standing expression, which are are central for language use. However, this behaviour was already displayed by the bidirectional approach without any regard for subphrases. Essentially, the subphraseness malus incentivizes patterns or phrases where

3-grams			4-grams		
phrase	score	$\#_{fg}$	phrase	score	$\#_{fg}$
Switzerland Global Enterprise	1.969314	5	Strich durch die Rechnung	2.049010	13
Musikantinnen und Musikanten	1.962437	15	St . Galler Tagblatt	1.956815	6
spot on news	1.955154	85	Leitindex Dow Jones Industrial	1.951936	5
nach wie vor	1.938053	309	Discounter Aldi und Lidl	1.949894	5
Golden State Warriors	1.931932	6	russische Präsident Wladimir Putin	1.910520	7
Höhen und Tiefen	1.928677	7	in Anspruch zu nehmen	1.887793	11
hin oder her	1.928631	11	liebe Leserinnen und Leser	1.884692	6
Swiss Performance Index	1.926714	7	spot on news AG	1.882229	80
Läuferinnen und Läufer	1.921929	18	türkische Präsident Recep Tayyip	1.873567	5
Bewohnerinnen und Bewohnern	1.909128	10	rund um den Globus	1.868047	10
zur Verfügung stehenden	1.907124	7	so gut wie möglich	1.867804	11
Mass aller Dinge	1.902209	12	this post on Instagram	1.867301	10
Schweizerinnen und Schweizern	1.893678	15	Präsidenten Recep Tayyip Erdogan	1.858150	6
Exklusiv für Abonnenten	1.882733	429	zur Verfügung zu stellen	1.852026	25
Teilnehmerinnen und Teilnehmer	1.881753	23	Russlands Präsident Wladimir Putin	1.851782	6
die Lupe genommen	1.880913	13	EuroStoxx 50 als Leitindex	1.850733	8
on news AG	1.875767	80	russischen Präsidenten Wladimir Putin	1.846006	9
Pierre Alain Schnegg	1.872137	17	rund um die Uhr	1.845506	42
Ein Beitrag geteilt	1.865863	11	gegen das Coronavirus geimpft	1.840897	5
Einwohnerinnen und Einwohnern	1.861119	21	es handle sich um	1.839217	9

Table 5.19: Top 20 3- and 4-grams for the bidirectional npmi with subphraseness bonus and a minimum count of 5

only the boundary tokens depend on each other, with as little disambiguating information content among the subphrases

5.2.2.3 Minimum Count Filter

I held out using a minimum count filter for as long as possible, but the approaches so far proved to favor rather infrequent phrases. The minimum count filter is problematic for two reasons. Firstly, it works directly on the word forms. This strongly punishes flexible phrases with differently inflected actualizations, and even more so phrases with slots such as "Alter von X Jahren" or "weder X noch Y". It moves the advantage to substantive idioms, away from formal idioms. Secondly, it introduces an arbitrary hyperparameter.

Nonetheless, the usage of a minimum count filter for phraseological unit identification can also be argued for. A minimum count filter ensures that the influence of the context C outside of the phrase is diminished even more, because the scores of the phrase in different contexts get averaged. The foreground corpus contained multiple duplicate sentences, but their influence only started showing in the rankings of 5- to 6-grams. I suspect this is due to the small size of the corpus, as with increasing phrase length there are fewer and fewer token sequences that occur with a certain minimum frequency.

Table 5.19 shows tri- and 4-grams for the bidirectional npmi with a subphraseness bonus and a minimum count of 5. Now there are almost no phrases that do not correlate with the human judgement regarding phraseness. Next to named entities, in the 4-grams sometimes even predicated is in “russische Präsident Wladimir Putin” or “Discounter Aldi und Lidl”, the trigrams display a few forms of gender-equitable noun phrases. But also, there are phrasemes such as the hendiadys “Höhen und Tiefen”, “nach wie vor” or “hin oder her”, as well as “zur Verfügung stehenden”, “Mass aller Dinge” or “die Lupe genommen”, usually used with the prefix “unter”. “unter die Lupe“, the fixed lexical contents of the phraseme “unter die Lupe nehmen”, resides in the top percentile on rank 350 of 44,359 trigrams meeting the filtering criterion. This isn't the first time in this qualitative discussion that the flexible verbs of a phrase appear to have a stronger dependence (or are more strongly depended on) than fixed prepositions.

The 4-grams also contain several substantive idioms such as “Strich durch die Rechnung”, “rund um den Globus” or “rund um die Uhr”, as well as encoding idioms as “in Anspruch zu nehmen”, “so gut wie möglich” and “es handle sich um”. Phrases that can be traced to duplicate sentences are the subphrases of “spot on news AG” or “Exklusiv für Abonnenten”, that appear in banners or source disclaimers.

5.2.2.4 Named Entity Filter

Finally, I want to present the results of the previous step, but after a heuristic named entity filter has been applied. In effect, I filter out all phrases that contain more than one capitalized letter. This of course does not suffice to effectively identify all named entities while not removing non-entities, but it still demonstrates how the remaining phrases, or rather, phrase parts, are of a high quality.

Once the minimum count filter is applied, the phrases of the lower tiers tend towards what Fillmore et al. [1988] called substantive idioms, i.e. idioms with fixed lexical contents But once more, boundary tokens of the idioms are not always part of the phrases. Furthermore, it appears that the upper tiers, i.e. 5- and 6-grams, present idioms that could be argued to be ‘key’ for the news domain. Several phrases show an act of citation or referral to an information source: “, die dieser Zeitung vorliegt”, “heisst es in der Mitteilung”, “heisst es im Strafbefehl”, “heisst es auf Anfrage .” and so forth. For these calculations, no algorithmic reference to a background is used anymore, as even the pretraining of SwissBERT was conducted on the news domain. However, I am well aware that it might as well be confirmation bias that draws me to this finding, and further experiments with multiple domains or subdomains would need to be conducted to examine this idea.

rank	2-grams	3-grams	4-grams
1	wünschen übrig	spot on news	rund um den Globus
2	geschweige denn	nach wie vor	in einem schlechten Zustand
3	Augen geführt	zur Verfügung stehenden	zur Verfügung zu stellen
4	fakultativen Referendum	hin oder her	gegen das Coronavirus geimpft
5	Anspruch genommen	hin und her	so gut wie möglich
6	willkommen heissen	fehl am Platz	rund um die Uhr
7	Erfüllung gegangen	auf sich aufmerksam	sich keiner Schuld bewusst
8	Beweis stellen	trägt dazu bei	in Anspruch zu nehmen
9	Revue passieren	ins Leben gerufen	. bis Do .
10	fehl am	alles andere als	in Verbindung zu setzen
11	zeugen davon	unter Beweis stellen	mit dem Schrecken davon
12	darum herum	die Lupe genommen	aus dem Toggenburg und
13	vom Fleck	in Kauf nehmen	sich auf den Standpunkt
14	eh und	zu Ende gegangen	nach wie vor unklar
15	Rolle gespielt	zur Verfügung steht	allem daran , dass
16	gehts zur	die Lupe nehmen	es handle sich um
17	darum gegangen	ausser Acht gelassen	aus dem Weg geräumt
18	Fallzahlen weltweit	zur Verfügung stellen	heisst es im Communiqué
19	Nikkei 225	ich mich fühle	zu einem späteren Zeitpunkt
20	am Hut	den Standpunkt gestellt	von der Hand zu
21	davon auszugehen	freuen wir uns	lange auf sich warten
22	japanische Yen	aus dem Ruder	auf das Coronavirus getestet
23	darauf zurückzuführen	um mich herum	von zu Hause aus
24	Gürtel enger	zur Verfügung stehen	liegen noch nicht vor
25	wählen lassen	schwarz auf weiss	mehr als 1500 Zeichen
26	Standpunkt gestellt	zur Welt gekommen	eine wichtige Rolle spielen
27	Plan gerufen	unter freiem Himmel	es handelt sich um
28	für Abonnenten	aus dem Handel	in Angriff zu nehmen
29	Wert darauf	in Kraft tritt	handelt sich um einen
30	wirft Fragen	zur Verfügung gestellt	ergeben nicht zwingend 100%

Table 5.20: Top 30 2-, 3- and 4-grams ranked by the bidirectional npmi with subphraseness bonus, a minimum count of 5 and a maximum of 1 capitalized letter (heuristic named entity filter)

rank	5-grams	6-grams
1	positiv auf das Coronavirus getestet	es in der Mitteilung weiter heisst
2	von der Hand zu weisen	, steht noch nicht fest .
3	geht es vor allem um	weiss , was ich kann und
4	, die nicht mehr als	auf das Coronavirus getestet wurden .
5	, wie ich mich fühle	: « Ich freue mich ,
6	Urteil ist noch nicht rechtskräftig	wie es in einer Mitteilung heisst
7	was ich kann und was	, heisst es im Strafbefehl .
8	habe ich noch nie erlebt	, ist noch nicht klar .
9	so viele wie noch nie	aus , dass es sich um
10	handelt es sich dabei um	, wie es weiter heisst .
11	nichts anderes übrig , als	machen können , werden gebeten ,
12	in den Sand zu stecken	und ergeben nicht zwingend 100% .
13	, die dieser Zeitung vorliegt	, wie die Gemeinde mitteilt .
14	konnte den Brand rasch löschen	wie es in der Mitteilung heisst
15	es in einer Mitteilung heisst	wie es in einer Mitteilung vom
16	heisst es in der Mitteilung	wie noch nie ins Ausland exportiert
17	so schwer verletzt , dass	Später stellte sich heraus , dass
18	Es handelt sich dabei um	informiert und verpasst keine News über
19	heisst es in einer Mitteilung	ist davon auszugehen , dass sich
20	heisst es in der Anklageschrift	freue mich sehr auf die neue
21	, ist noch unklar .	, heisst es im Bericht .
22	, wie die Gemeinde mitteilt	, heisst es im Communiqué .
23	», wie es heisst .	wie er in einer Mitteilung schreibt
24	es in der Mitteilung heisst	stellte sich heraus , dass es
25	vor , Texte zu kürzen	, den in den vergangenen Jahren
26	mehr als im Jahr davor	ist selber an der Börse aktiv
27	, wenn immer möglich ,	« Es geht darum , dass
28	heisst es in der Studie	die sich auf aktuelle Artikel beziehen
29	Spekulationen und alles , was	dass ihr die Kraft fehle ,
30	seit mehr als zehn Jahren	, heisst es auf Anfrage .

Table 5.21: Top 30 5 and 6-grams ranked by the bidirectional npmi with subphraseness bonus, a minimum count of 5 and a maximum of 1 capitalized letter (heuristic named entity filter

5.2.2.5 Discussion

The biggest increase of quality I presented for phrase identification by probing an MLM comes from using a bidirectional measure and from introducing a minimum count filter. The concept of phraseness, i.e. how strongly words are associated, is demonstrably within reach by tracking the probabilities of tokens depending on how their context is masked. However, the raw output of the measures introduced is not directly usable as is. This becomes clear when considering the several phrases that are cut off, either when the fixed lexical contents of the phrase don't appear in boundary positions (e.g. "zwischen X und Y") or when, for reasons not entirely clear to me, when fixed prefixes such as "unter" for "unter die Lupe nehmen" only appear late in the ranking. Some form of post-processing needs to be undertaken, I resorted to different filters, but phrase extension algorithms and fuzzy matching strategies of the same phrases with different fillers come to mind.

Another point of contention are sentence duplicates. While the presented approaches are not as sensitive towards duplicates as traditional corpus linguistic association measures, this is still a point that could easily be remedied. One might also test an incentive for a high variability in the context of the different occurences of the phrases, analoguous to YAKE!, although differently motivated.

Finally, the role of the corpus and the role of the language model needs to be called into question. I approached these final phraseness experiments with a concept-, code- and resource-base from keyphrase extraction. I used the finetuned model and inferred the probabilities on the sentences of the foreground corpus. The identified phrases did not appear to be particularly connected to the January of 2023, at most they seemed indicative of media discourse. Therefore, let me call into question: Was I identifying the phrases of the foreground corpus, or was I identifying dependencies of the language model, and simply using the foreground corpus as a seed bank? I am tending towards the latter. The finetuning and the frequency filter of course moves the results towards the distributions of the foreground corpus, but the essence of the approaches lies in the language model.

6 Conclusion

This thesis attempted to translate an old Keyphrase Extraction approach from the Paper "A language model approach to keyphrase extraction" by Tomokiyo and Hurst [2003] into modern NLP technology, namely with the MLM SwissBERT. It failed in doing so. The probabilistic properties of the MLM and its subword tokenization deviate too strongly from the n-gram language models of the original paper, for the proposed pointwise Kullback-Leibler divergence to yield informative and understandable keyphrases. However, I believe it can be considered a productive failure.

In preparing the theoretical foundations for the evaluation and the discussion of the experimental results, several important concepts to keyphrases have been identified and made explicit. By comparing different methodologies and frameworks to keyphrases from applied NLP research, corpus linguistics and linguistics in general, implicit differences and their consequences were able to be named.

On a practical level, the attempted translation of the older approach hinted at the capacity of probing MLMs to identify phraseologisms or idiomatic expressions. Further experiments confirmed this capacity. When using the normalized pointwise mutual information (npmi) to compare the probabilities of a token once more and more of its context is masked, we are able to calculate how strongly tokens depend on each other. The qualitative discussion of the results corroborated that this approach effectively models idiomaticity in a broad sense, conceptualized by Fillmore et al. [1988] and the modern pattern or construction grammar.

6.1 Answers to the Research Questions

How can the Keyphrase Extraction approach by Tomokiyo and Hurst [2003] be translated to MLMs?

The main idea behind Tomokiyo and Hurst [2003] is to compare the probabilities of phrases when considering the empirical distribution and when assuming independence between the phrase's constituents. This idea can be translated to MLMs by extracting probabilities for a single token x with differently masked contexts. It allows us to effectively measure the amount of information that was carried by the masked tokens for this token x, especially when averaging over different contexts.

How does the adapted approach perform for Keyphrase Extraction, and what explains this performance?

The translation performs poorly on the task of Keyphrase Extraction. This is partly due to a scaling mismatch of the scoring components, which is especially pronounced with MLM. On the other hand, the used measure of the pointwise Kullback-Leibler divergence incentivizes tokens that are highly improbable in the background corpus. This leads to keyphrases containing rare tokens.

How can the methodology developed for Keyphrase Extraction be used outside of this context?

Experiments outside of Keyphrase Extraction have shown how probing an MLM with differently masked sentences can lead to the identification of idiomatic structures. This is noteworthy as even very formalized, syntactical idioms such as “weder X noch Y” or “zwischen X und Y” get identified, next to a plethora of substantive idioms like “Dorn im Auge” or “eine wichtige Rolle spielen”. However, I can not point to a single variant of the approach that stands out among all others, and while the results are promising, postprocessing would be needed to form a full-on idiom identification framework.

6.2 Further Research

Much remains to be explored. The approach I presented is heavily biased by the subword tokenization, as I effectively could only consider subword tokens that constituted whole words. Furthermore, the approach frequently showed the same pattern where boundary tokens of known idioms are not included in the high ranking phrases. Further research might be concerned with developing postprocessing techniques such as phrase extension, using this method as a seed generator. Next thing to mention is how this approach is rather resource intensive. Even very short sentences need to be inferred many different times with different masking patterns. Salazar et al. [2020] and Kauf and Ivanova [2023] hint at different possibilities to mitigate this, but in the context of perplexity estimation.

For the total thesis, I regarded a whole foreground corpus, motivated by the initial goal of Keyphrase Extraction. However, nothing prevents the defined measures to be applied to single documents and sentences, where they might be used for chunking and for data-driven, shallow parsing. Furthermore, let me reiterate again on the remark from last chapter's conclusion. Are there possibilities to directly query an MLM's representation of phraseness and idiomaticity without resorting to using the sentences of a specific domain as seeds? Maybe the search space can be controlled in other ways.

Especially the phraseness results without a minimum count constraint demonstrated how formal idioms, patternized syntactical constructions, are able to be identified. Future research ought to investigate this further, developing methods to account for and unify the variable lexical fillers of the phrase's slots. This might include using a more refined, more resourceful masking strategy than the sliding window of incremental size that I employed. Additionally, for all of the phraseness work I undertook, I focused on how the probability decreases once context is masked, and only hinted at the possibility of investigating increasing probabilities. I suspect that this other focus might be very fruitful for phraseological research to evaluate semantic constrictions on phrase slots. Even further, this could prove to be a viable way of assessing biases of MLMs.

References

R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Santiago, Chile, 1994.

A. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, Jan. 2003. doi: 10.1016/S0306-4573(02/00021-3. URL `https://www.sciencedirect.com/science/article/pii/S0306457302000213`.

R. C. Atkinson. Mnemotechnics in second-language learning. *American Psychologist*, 30(8):821–828, 1975. doi: 10.1037/h0077029.

P. Baker. Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, 32(4):346–359, Dec. 2004. doi: 10.1177/0075424204269894. URL `http://journals.sagepub.com/doi/10.1177/0075424204269894`.

K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings, Sept. 2018. URL `http://arxiv.org/abs/1801.04470`.

L. Boltzmann. *Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen: vorgelegt in der Sitzung am 10. October 1872.* k. und k. Hof- und Staatsdr., 1872.

F. Boudin. TALN Archives : a digital archive of French research articles in Natural Language Processing (TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue) [in French]. In E. Morin and Y. Estève, editors, *Proceedings of TALN 2013 (Volume 2: Short Papers)*, pages 507–514, Les Sables d'Olonne, France, June 2013. ATALA. URL `https://aclanthology.org/F13-2001`.

F. Boudin. Unsupervised Keyphrase Extraction with Multipartite Graphs. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 2 (Short Papers), pages 667–672, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2105. URL https://aclanthology.org/N18-2105.

A. Bougouin, F. Boudin, and B. Daille. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In R. Mitkov and J. C. Park, editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan, Oct. 2013. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I13-1062.

V. Brezina. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press, 1 edition, Sept. 2018. doi: 10.1017/9781316410899. URL https://www.cambridge.org/core/product/identifier/9781316410899/type/book.

N. Bubenhofer. 4. Kollokationen, n-Gramme, Mehrworteinheiten. In K. S. Roth, M. Wengeler, and A. Ziem, editors, *Handbuch Sprache in Politik und Gesellschaft*, pages 69–93. De Gruyter, Apr. 2017. doi: 10.1515/9783110296310-004. URL https://www.degruyter.com/document/doi/10.1515/9783110296310-004/html.

R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt. YAKE! Collection-Independent Automatic Keyword Extractor. volume 10772, pages 806–810, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-319-76941-7_80. URL http://link.springer.com/10.1007/978-3-319-76941-7_80.

R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2019. doi: 10.1016/j.ins.2019.09.013. URL https://www.sciencedirect.com/science/article/pii/S0020025519308588.

C. Caragea, F. A. Bulgarov, A. Godea, and S. Das Gollapalli. Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1150. URL https://aclanthology.org/D14-1150.

J. B. Carroll. An Alternative to Juilland's Usage Coefficient for Lexical Frequencies1. *ETS Research Bulletin Series*, 1970(2):i–15, 1970. ISSN 2333-8504. doi: 10.1002/j.2333-8504.1970.tb00778.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1970.tb00778.x.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, Oct. 1999. doi: 10.1006/csla.1999.0128. URL https://linkinghub.elsevier.com/retrieve/pii/S0885230899901286.

C. Cole. Shannon revisited: Information in terms of uncertainty. *Journal of the American Society for Information Science*, 44(4):204–211, 1993. doi: 10.1002/(SICI/1097-4571(199305/44:4<204::AID-ASI3>3.0.CO;2-4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199305%2944%3A4%3C204%3A%3AAID-ASI3%3E3.0.CO%3B2-4.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, Apr. 2020. URL http://arxiv.org/abs/1911.02116.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1 edition, Sept. 2005. ISBN 978-0-471-24195-9 978-0-471-74882-3. doi: 10.1002/047174882X. URL https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X.

W. Croft. Construction Grammar. In D. Geeraerts and H. Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*, pages 464–508. Oxford University Press, June 2010. doi: 10.1093/oxfordhb/9780199738632.013.0018. URL https://doi.org/10.1093/oxfordhb/9780199738632.013.0018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805.

H. Ding and X. Luo. AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1928, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.146. URL https://aclanthology.org/2021.emnlp-main.146.

T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993. URL https://aclanthology.org/J93-1003.

J. Egbert and D. Biber. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104, Apr. 2019. doi: 10.3366/cor.2019.0162. URL https://www.euppublishing.com/doi/abs/10.3366/cor.2019.0162.

S. Evert, N. Dykes, and J. Peters. A quantitative evaluation of keyword measures for corpus-based discourse analysis. In *Proceedings of Corpora and Discourse Conference*, Lancaster, 2018.

R. M. Fano. *Transmission of Information: A Statistical Theory of Communications.* MIT press, 1961. URL https://www.jstor.org/stable/10.2307/2982638?origin=crossref.

C. J. Fillmore, P. Kay, and M. C. O'Connor. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Langugae*, 64(3):501–538, 1988.

C. Gabrielatos. Keyness analysis. In E. Friginal and J. A. Hardy, editors, *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, pages 225–258. 2018. Publisher: Routledge London.

Y. Gallina, F. Boudin, and B. Daille. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents, Nov. 2019. URL http://arxiv.org/abs/1911.12559.

S. D. Gollapalli and C. Caragea. Extracting Keyphrases from Research Papers Using Citation Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28 (1), June 2014. doi: 10.1609/aaai.v28i1.8946. URL https://ojs.aaai.org/index.php/AAAI/article/view/8946. Number: 1.

W. Grabe. Vocabulary and reading comprehension. In *Reading in a Second Language: Moving from Theory to Practice*, Cambridge Applied Linguistics, pages 265–286. Cambridge University Press, Cambridge, 2008. doi: 10.1017/CBO9781139150484.018. URL https://www.cambridge.org/core/books/reading-in-a-second-language/vocabulary-and-reading-comprehension/76B908FC787A6BB48C9FE19E987D0FA9.

H. P. Grice. Logic and conversation. In P. Cole, editor, *Syntax and semantics 3: Speech acts*, pages 41–58. Academic Press, New York, 1975.

S. T. Gries. Phraseology and linguistic theory. In S. Granger and F. Meunier, editors, *Phraseology: an interdisciplinary perspective.* John Benjamins Publishing, Amsterdam ; Philadelphia, 2008.

S. T. Gries. Analyzing Dispersion. In M. Paquot and S. T. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pages 99–118. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-46216-1_5. URL https://link.springer.com/10.1007/978-3-030-46216-1_5.

S. T. Gries. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2):1–33, Jan. 2021. doi: 10.32714/ricl.09.02.02. URL https://ricl.aelinco.es/index.php/ricl/article/view/150.

X. Gu, Z. Wang, Z. Bi, Y. Meng, L. Liu, J. Han, and J. Shang. UCPhrase: Unsupervised Context-aware Quality Phrase Tagging. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 478–486, Aug. 2021. doi: 10.1145/3447548.3467397. URL http://arxiv.org/abs/2105.14078.

K. S. Hasan and V. Ng. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In C.-R. Huang and D. Jurafsky, editors, *Coling 2010: Posters*, pages 365–373, Beijing, China, Aug. 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-2042.

K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, 2014. URL https://aclanthology.org/P14-1119.pdf.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-Efficient Transfer Learning for NLP, June 2019. URL http://arxiv.org/abs/1902.00751.

HuggingFace. Perplexity of fixed-length models. URL https://huggingface.co/docs/transformers/perplexity.

S. Hunston and G. Francis. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* John Benjamins Publishing, Amsterdam, 2000.

A. Juilland and E. Chang-Rodriguez. Frequency Dictionary of Spanish Words. In *Frequency Dictionary of Spanish Words.* De Gruyter Mouton, Berlin, 1964. doi: 10.1515/9783112415467. URL https://www.degruyter.com/document/doi/10.1515/9783112415467/html.

D. Jurafsky and J. H. Martin. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Youcanprint, 2023. ISBN 9791221476842. URL https://web.stanford.edu/~jurafsky/slp3/.

C. Kauf and A. Ivanova. A Better Way to Do Masked Language Model Scoring, May 2023. URL http://arxiv.org/abs/2305.10588.

T. Kew, M. Kostrzewa, and S. Ebling. 20 Minuten: A Multi-task News Summarisation Dataset for German. June 2023. doi: 10.5167/UZH-234387. URL https://www.zora.uzh.ch/id/eprint/234387.

A. Kilgarriff. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155, June 1997. doi: 10.1093/ijl/10.2.135. URL https://academic.oup.com/ijl/article-lookup/doi/10.1093/ijl/10.2.135.

A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, and X. Bai. PromptRank: Unsupervised Keyphrase Extraction Using Prompt. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9788–9801, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.545. URL https://aclanthology.org/2023.acl-long.545.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Aug. 2018. URL http://arxiv.org/abs/1808.06226.

G. Lample and A. Conneau. Cross-lingual Language Model Pretraining, Jan. 2019. URL http://arxiv.org/abs/1901.07291.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692.

Z. Liu, P. Li, Y. Zheng, and M. Sun. Clustering to Find Exemplar Terms for Keyphrase Extraction. In P. Koehn and R. Mihalcea, editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 257–266, Singapore, Aug. 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1027.

Z. Liu, W. Huang, Y. Zheng, and M. Sun. Automatic Keyphrase Extraction via Topic Decomposition. In H. Li and L. Màrquez, editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376, Cambridge, MA, Oct. 2010. Association for Computational Linguistics. URL https://aclanthology.org/D10-1036.

D. J. MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

D. Mahata, J. Kuriakose, R. R. Shah, and R. Zimmermann. Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In

M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2100. URL https://aclanthology.org/N18-2100.

R. Mihalcea and P. Tarau. TextRank: Bringing Order into Text. In D. Lin and D. Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3252.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. Nov. 1999. URL https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking-%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556.

E. Papagiannopoulou and G. Tsoumakas. Unsupervised Keyphrase Extraction from Scientific Publications. volume 13451, pages 215–229, Cham, 2018. Springer Nature Switzerland. doi: 10.1007/978-3-031-24337-0_16. URL https://link.springer.com/10.1007/978-3-031-24337-0_16.

E. Papagiannopoulou and G. Tsoumakas. A review of keyphrase extraction. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1339, 2020. doi: 10.1002/widm.1339. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1339.

M. Paquot. *Academic vocabulary in learner writing: From extraction to analysis.* Corpus and discourse. Bloomsbury Publishing, 2010. URL https://books.google.ch/books?id=gVU7BAAAQBAJ.

R. J. Passonneau and I. Mani. Evaluation. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 435–453. Oxford University Press, 2014. ISBN 978-0-19-957369-1. doi: 10.1093/oxfordhb/9780199573691.013.014. URL https://doi.org/10.1093/oxfordhb/9780199573691.013.014.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations, Mar. 2018. URL http://arxiv.org/abs/1802.05365.

M. Peterson and E. Salvagio. Figure-ground perception. *Scholarpedia*, 5(4):4320, Apr. 2010. doi: 10.4249/scholarpedia.4320. URL http://www.scholarpedia.org/article/Figure-ground_perception.

J. Pfeiffer, N. Goyal, X. Lin, X. Li, J. Cross, S. Riedel, and M. Artetxe. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255.

T. R. C. Read and N. A. C. Cressie. Historical Perspective: Pearson's X2 and the Loglikelihood Ratio Statistic G2. In T. R. C. Read and N. A. C. Cressie, editors, *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer Series in Statistics, pages 133–153. Springer, New York, NY, 1988. URL https://doi.org/10.1007/978-1-4612-4578-0_9.

S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, Oct. 2004. doi: 10.1108/00220410410560582. URL https://www.emerald.com/insight/content/doi/10.1108/00220410410560582/full/html.

J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, 2020. doi: 10.18653/v1/2020.acl-main.240. URL http://arxiv.org/abs/1910.14659.

M. Scott. PC analysis of key words — And key key words. *System*, 25(2):233–245, June 1997. doi: 10.1016/S0346-251X(97/00011-0. URL https://www.sciencedirect.com/science/article/pii/S0346251X97000110.

M. Scott. Word Smith Tools version 3.0, 1998. URL https://lexically.net/wordsmith/version3/manual.pdf.

C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948. URL https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

W. Shi, W. Zheng, J. X. Yu, H. Cheng, and L. Zou. Keyphrase Extraction Using Knowledge Graphs. *Data Science and Engineering*, 2(4):275–288, Dec. 2017. doi: 10.1007/s41019-017-0055-z. URL https://doi.org/10.1007/s41019-017-0055-z.

M. Song, Y. Feng, and L. Jing. A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models. In *Findings of the Association for Computational*

Linguistics: EACL 2023, pages 2153–2164, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.161.

K. Sparck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21, Jan. 1972. ISSN 0022-0418. doi: 10.1108/eb026526. URL https://www.emerald.com/insight/content/doi/10.1108/eb026526/full/html.

A. Stefanowitsch. *Corpus linguistics. A guide to the methodology.* Number 7 in Textbooks in Language Sciences. Language Science Press, Berlin, 2020. URL http://langsci-press.org/catalog/book/148.

L. Sterckx, T. Demeester, J. Deleu, and C. Develder. Topical Word Importance for Fast Keyphrase Extraction. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 121–122, New York, NY, USA, May 2015. Association for Computing Machinery. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2742730. URL https://dl.acm.org/doi/10.1145/2740908.2742730.

M. Stubbs. Three concepts of keywords. In *Proceedings of Keyness in texts*, pages 21–42, 2010.

L. Sönning. Evaluation of keyness metrics: performance and reliability. *Corpus Linguistics and Linguistic Theory*, Apr. 2023. ISSN 1613-7035. doi: 10.1515/cllt-2022-0116. URL https://www.degruyter.com/document/doi/10.1515/cllt-2022-0116/html. Publisher: De Gruyter Mouton.

W. L. Taylor. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433, Sept. 1953. doi: 10.1177/107769905303000401. URL http://journals.sagepub.com/doi/10.1177/107769905303000401.

T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions analysis, acquisition and treatment*, volume 18, pages 33–40. Association for Computational Linguistics, 2003. doi: 10.3115/1119282.1119287. URL http://portal.acm.org/citation.cfm?doid=1119282.1119287.

P. Turney. Learning to Extract Keyphrases from Text. Departmental Technical Report, National Research Council of Canada, 1999. URL https://web-archive.southampton.ac.uk/cogprints.org/1802/index.html.

J. Vamvas and R. Sennrich. Towards Unsupervised Recognition of Token-level Semantic Differences in Related Documents, Oct. 2023. URL http://arxiv.org/abs/2305.13303.

J. Vamvas, J. Graën, and R. Sennrich. SwissBERT: The Multilingual Language Model for Switzerland, Mar. 2023. URL http://arxiv.org/abs/2303.13310. arXiv:2303.13310 [cs].

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need, 2017. URL http://arxiv.org/abs/1706.03762.

X. Wan and J. Xiao. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of AAAI 2008*, volume 8, pages 855–860, July 2008.

A. Wang and K. Cho. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model, Apr. 2019. URL http://arxiv.org/abs/1902.04094.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, Feb. 2019. URL http://arxiv.org/abs/1804.07461.

R. Wang, W. Liu, and C. Mcdonald. Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors. volume 39, pages 1–8, 2014, 2015. URL https://www.semanticscholar.org/paper/Corpus-independent-Generic-Keyphrase-Extraction-Wang-Liu/bd3794c777af5ba363abae5708050ea78ecc97e2.

R. Williams. *Keywords: A Vocabulary of Culture and Society.* Oxford University Press, USA, May 1985.

Y. Yu and V. Ng. WikiRank: Improving Keyphrase Extraction Based on Background Knowledge, Mar. 2018. URL http://arxiv.org/abs/1803.09000. arXiv:1803.09000 [cs].

T. Zesch and I. Gurevych. Approximate Matching for Evaluating Keyphrase Extraction. In G. Angelova and R. Mitkov, editors, *Proceedings of the International Conference RANLP-2009*, pages 484–489, Borovets, Bulgaria, Sept. 2009. Association for Computational Linguistics. URL https://aclanthology.org/R09-1086.

L. Zhang, Q. Chen, W. Wang, C. Deng, S. Zhang, B. Li, W. Wang, and X. Cao. MDERank: A Masked Document Embedding Rank Approach for Unsupervised Keyphrase Extraction, Feb. 2023. URL http://arxiv.org/abs/2110.06651.

University of Zurich^UZH

Department of Computational Linguistics

Declaration of Independent Authorship

Original work

I expressly declare that the written work I submitted to the University of Zurich in the spring/autumn semester of 2023 with the title

(Key) Phrase Extraction with Masked Language Models and Information Theory

is an original work written by myself, in my own words, and without unauthorized assistance. If it is a work by several authors, I confirm that the relevant parts of the work are correctly and clearly marked and can be clearly assigned to the respective author.

I also confirm that the work has not been submitted in whole or in part to receive credit for another module at the University of Zurich or another educational institution, nor will it be submitted in the future.

Use of sources

I expressly declare that I have identified all references to external sources (including tables, graphics etc.) contained in the above work as such. In particular, I confirm that, without exception and to the best of my knowledge, I have indicated the authorship both for verbatim statements (citations) and for statements by other authors reproduced in my own words (paraphrases).

Use of text generation models

I expressly declare that I have not only identified existing external sources, but also any automatically generated text that is contained in the above work. I have used the same citation style as if the text had been generated by a human to indicate the source of the automatically generated text. If the contribution of text generation models cannot be linked to specific text passages (see the associated guidelines), I have included a chapter describing the contributions of the text generation model. I acknowledge that no explicit citation is necessary where text generation models are merely used correctively (to improve grammar or idiomaticity of my own words).

Sanctions

I acknowledge that a thesis that is used to acquire credit and proves to be plagiarism with the meaning of the document *Erläuterung des Begriffs „Plagiat“* leads to a grade deduction in minor cases, a grade 1 (one) in more severe cases, without the possibility of revision, and in very severe cases can have the corresponding legal and disciplinary consequences according to §§ 7ff of the “Disziplinarordnung der Universität Zürich” and § 36 of the “Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich”.

I confirm with my signature that this information is correct:

Name: Bodenmann First Name: Niclas

Matriculation number: 17-700-436

Date: 31.12.2023 Signature: