Universität Zürich[UZH]

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

Domänenspezifische neuronale maschinelle Übersetzung für finanzwirtschaftliche Texte

Verfasserin: Mara Bertamini
MA Multilinguale Textanalyse
Matrikel-Nr: 11-919-347
Sonnmattweg 1
5620 Bremgarten
maratiziana.bertamini@uzh.ch

Referent: Prof. Dr. Martin Volk
Betreuer: Samuel Läubli
Institut für Computerlinguistik

Abgabedatum: 18.06.2018

Zusammenfassung

Diese Arbeit behandelt das Training eines neuronalen Übersetzungssystems (EN → DE) mit generischen Daten, welches in einem zweiten Schritt an die Domäne von firmenspezifischen, finanzwirtschaftlichen Texten adaptiert wird. Ziel dieses Experimentiersettings ist es, anhand des Toolkits Nematus ein neuronales *State-of-the-Art*-Übersetzungssystem an eine neue, in der Literatur noch wenig diskutierte Textsorte anzupassen. Eine differenzierte Analyse – unter Berücksichtigung der sowohl quantitativen als auch qualitativen Dimension – soll Aufschluss über die Performanz des Domänen-adaptierten Systems geben und damit neue Denkanstösse in Punkto Datenmenge, Qualität und Evaluation liefern. Gesamthaft liegt der Fokus der Arbeit auf dem Aufbau und der Vorverarbeitung des domänenspezifischen Korpus' sowie der qualitativen Evaluation der Ergebnisse.

Abstract

This thesis deals with the training of a neural translation system (EN → DE) with generic data, which in a second step is adapted to the domain of company-specific financial texts. The aim of this experimental setting is to adapt a neural *State-of-the-Art* translation system to a new type of text that is still little discussed in the literature using the Nematus toolkit. A differentiated analysis – taking into account the quantitative as well as qualitative dimension – should shed light on the performance of the domain-adapted system and thus provide new food for thought in terms of data volume, quality and evaluation. Overall, this thesis focuses on the development of the domain-specific corpus and the qualitative evaluation of the results.

Danksagung

Mein Dank gilt einer Reihe von Personen, die mich bei der Durchführung dieses Masterprojekts unterstützt haben:

Martin, für die unkomplizierte Zusammenarbeit und die Ermutigung, bei dieser Arbeit meine Komfortzone zu verlassen; Samuel, für die uneitle Unterstützung, die guten Inputs und alle möglichen Gratis-Tipps und Tricks. Eveline, Steffi und Bruno, für eure Flexibilität und euer Verständnis, dass ich mich des Masterprojekts wegen etwas seltener bei der Arbeit blicken liess.

Ein grosses Dankeschön geht ebenso an meine Familie, die mich (sowohl finanziell als auch moralisch) stets ermutigt hat, meine universitäre Ausbildung voranzutreiben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

1.1 Motivation

Meine persönliche Motivation für diese Arbeit gründet auf zwei Hauptpfeilern: Einerseits fasziniert mich Übersetzung – ob menschlich oder maschinell, bzw. in Kombination miteinander – generell, andererseits habe ich ein grosses Interesse an sprachtechnologischen Anwendungen für die Industrie, genauer gesagt für Firmen im Finanzsektor. Während meiner nun 2.5-jährigen Tätigkeit beim grössten Schweizer Lebensversicherungskonzern Swiss Life hatte ich immer wieder das Privileg, Einblicke in diverse sowohl zahlen- als auch textgetriebene Bereiche zu erhalten und darüber nachzudenken, in welchem Unternehmensbereich sich Sprachtechnologie einsetzen liesse und v.a. wie sich diese Integration schlussendlich auf die Effizienz und Produktivität der Firma auswirken könnte.

Der Finanzsektor, bestehend aus dem Banken- und Versicherungssektor, ist insofern eine vielversprechende Branche bzgl. Sprachtechnologie, als dass Compliance und Vertraulichkeit grossgeschrieben werden, d.h. es sollen möglichst keine Daten ausserhalb des Unternehmens gelangen, was nach auf die Firma zugeschnittenen Lösungen ruft. Dass zuweilen dennoch interne Daten nach aussen gelangen, wenn beispielsweise ein Textsegment durch GoogleTranslate geschickt wird, ist – so behaupte ich – vielen Mitarbeitenden eines Unternehmens nicht bewusst ist. Mögen sich die Schäden für Privatpersonen in Grenzen halten, so möchte man als Firma vermeiden, dass interne, vertrauliche Daten mittels der Verwendung von frei verfügbaren Online-Diensten bei sog. Tech-Giganten landen, die ebendiese Online-Dienste zur Verfügung stellen. Ein internes maschinelles Übersetzungssystem könnte gewährleisten, dass interne Daten nicht an eigentlich unbefugte Dritte gelangen; zudem wäre die Qualität der Übersetzungen sichergestellt, sofern das interne System auf domänenspezifischen – und nicht auf generischen oder gar gemischten, domänenfremden – Daten trainiert worden ist.

Nebst den eben beschriebenen regulatorischen Gründen spielt natürlich auch die Menge an zu übersetzenden Texten eine wichtige Rolle: Die Anzahl Texte nimmt

stetig zu, sodass (firmeninterne) Übersetzer mehr oder minder dauerhaft ausgelastet sind; dementsprechend dauern die Übersetzungsprozesse – je nach Dringlichkeit des Auftrags – aus Sicht der Auftraggeber (zu) lange. Maschinelle Übersetzung könnte hier insofern Abhilfe schaffen, als dass sprachlich regelmässige, strukturierte Texte (wie z.B. Weisungen, Produktbeschreibungen oder IT-Dokumente) mehr oder weniger gänzlich maschinell übersetzt werden könnten; dabei bliebe mehr Zeit für die Übersetzung von redaktionellen Texten (die, so wage ich zu behaupten, für professionelle Übersetzer sowieso spannender sind).

Die Integration von maschineller Übersetzung in den zuweilen noch menschlichen Übersetzungsprozess ist ein anderes Thema, das zu bearbeiten den Rahmen dieser Arbeit sprengen würde. Nichtsdestotrotz scheint es mir wichtig hervorzuheben, dass sowohl die Akzeptanz von maschineller Übersetzung als auch ein gewisser Wissensstand bzgl. der Funktionsweise derselben von Seiten der professionellen Übersetzer kritische Erfolgsfaktoren für eine erfolgreiche Implementierung sind. Es reicht nicht, ein performantes maschinelles Übersetzungssystem zu konstruieren, ohne diejenigen Personen, die schlussendlich mit dem maschinellen Output arbeiten, für das Thema zu sensibilisieren und angemessene Schulungen durchzuführen.

Versicherungskonzerne, wie Swiss Life einer ist, sind aber nicht nur aufgrund der beschriebenen regulatorischen Compliance und der optimalen Datenlage potentielle Kunden für sprachtechnologische Anwendungen wie z.B. maschinelle Übersetzung: In der Finanzbranche als kapitalstärkster Sektor (in der Schweiz hat der Finanzsektor bspw. einen Anteil von rund 10% am gesamten Wirtschaftswachstum[1]) sind schlicht die liquiden Mittel vorhanden, die schlussendlich unerlässlich sind, um firmenspezifische, sprachtechnologische Lösungen überhaupt erst entwickeln lassen zu können.

1.2 Zur Domänenadaption

Unter Domänenadaption subsummiert Sennrich (2013) in seiner Dissertation "[...] all methods that try to make better use of the part of training data that is more similar, and thus more relevant, to the text that is being translated". Angesichts der im vorherigen Unterkapitel angepriesenen textuellen Datenfülle im Finanzsektor mag es durchaus unnötig erscheinen, Domänenadaption für diese Domäne überhaupt prüfen zu wollen, da ja genügend relevante, i.e. *in-domain*-Trainingsdaten vorhanden sein sollten und darum nicht zwingend auf einen ergänzenden, grösseren Datensatz

[1] `http://www.swissbanking.org/de/finanzplatz/finanzplatz-in-zahlen/der-schweizer-finanzsektor`

an *out-of-domain*-Material zurückgegriffen werden müsste.

Fakt ist aber, dass parallele, finanzwirtschaftliche Texte einerseits nur relativ mühsam (gratis und ohne juristische Folgen) zu erhalten sind (gemäss CLARIN[2] gibt es keine frei verfügbaren, rein Finanzsektor-spezifische parallelen Korpora, abgesehen von Textsammlungen aus Finanzmagazinen); die (manuelle) Aufbereitung von Rohdaten – wenn man denn welche hat – ist aufwändig und zeitintensiv, wie in Kapitel 3 genauer beschrieben wird. Chen et al. (2017) halten zudem fest, dass "Domain adaptation for NMT is still a new research area, with only a small number of relevant publications". Daher ist es m.E. umso wichtiger, die Methode der Domänenadaption an sich, aber auch speziell für Daten aus dem Finanzsektor zu analysieren, denn nebst grossen, prominenten Akteuren in dieser Branche gibt es auch eine Vielzahl an kleineren Institutionen, welche durchaus ein Interesse an maschineller Übersetzung haben könnten, allerdings schlicht nicht über die für neuronale Systeme nötige (parallele) Datenmenge verfügen.

Des Weiteren bleibt zu evaluieren, inwieweit anhand relativ simpler und schneller Adaptionsmethoden eine zufriedenstellende Übersetzungsqualität erreicht werden kann. Domänenadaption bildet, wie Khayrallah et al. (2017) beschreiben, eine grosse Herausforderung für neuronale maschinelle Übersetzungssysteme, da diese Art von Architekturen einen grossen Trainingsdatensatz benötigen und darum der Output in ressourcenarmen sowie Domänenadaptions-Szenarien (im Vergleich zu phrasenbasierten Systemen unter denselben Konditionen) relativ schlecht ist: Der Output ist zwar sehr fliessend, enthält aber des öfteren Wörter, die semantisch nicht dem Input entsprechen. Die relevantesten Publikationen zu Domänenadaption in einem neuralen Setting, welche in Kapitel 2 noch genauer erläutert werden, lieferten u.a. Luong und Manning (2015) mit *Finetuning*, wo ein bestehendes *out-of-domain*-Modell für einige Epochen auf *in-domain*-Daten trainiert wird sowie Miceli Barone et al. (2017), welche in Kombination mit *Finetuning* die Methode *Transfer Learning* vorschlagen, um der Gefahr von *Overfitting* entgegenwirken zu können; Freitag und Al-Onaizan (2016) mit einem *Ensemble*-Modell, bei welchem das *out-of-domain*-Modell mit dem neu trainierten *in-domain*-Modell kombiniert wird; Chu et al. (2017) mit *mixed Finetuning*, wo die Ansätze *Finetuning* und *Multi-Domain*-NMT miteinander kombiniert werden; sowie Chen et al. (2017) mit einem sog. *Cost Weighting*-Ansatz, in welchem ein *Domain Classifier* angewendet wird, welcher für jeden Satz im Trainingskorpus entscheidet, ob es sich um *in-domain*- oder *out-of-domain*-Material handelt, was das Modell zu Gunsten der *in-domain*-Daten beeinflusst.

Der Anspruch dieser Arbeit ist es darum, diese oben beschriebene Forschungslücke

[2] `https://www.clarin.eu/news/parallel-corpora-overview`

– wenn auch nur zu einem kleinen Teil – mit neuen Erkenntnissen zu füllen und textsortenspezifische Herausforderungen ans Licht zu bringen.

1.3 Forschungsfragen

Die primären Forschungsfragen, auf denen diese Arbeit gründet, lauten:

1. Wie gut funktioniert Domänenadaption (anhand eines eigens kreierten Korpus' à 37'000 parallelen Sätzen) für finanzwirtschaftliche Texte?
2. Durch welche (linguistischen) Charakteristika zeichnet sich die finanzwirtschaftliche Domäne aus?
3. Wie konsistent werden diese in den maschinellen deutschen Übersetzungen beibehalten?

1.4 Aufbau der Arbeit

Der Aufbau der Arbeit lautet wie folgt: Kapitel 2 informiert über den aktuellen Forschungsstand; Kapitel 3 beinhaltet eine Beschreibung der verwendeten Daten sowie eine detaillierte Übersicht zu allen Teilen des Pre-Processings; Kapitel 4 ist dem Training des neuronalen Netzes gewidmet; Tuning wird in Kapitel 5 behandelt. Die Evaluation und Fehleranalyse sind in Kapitel 6 enthalten, während Kapitel 7 die Arbeit mit einem Fazit abrundet.

2 Stand der Forschung

2.1 Neuronale maschinelle Übersetzung

2.1.1 Idee / Konzept

Das Interesse an einem neuen Paradigma in der maschinellen Übersetzung ist in den letzten Jahren rasant gestiegen: Neuronale maschinelle Übersetzungsmodelle versprechen eine bessere Ausnutzung von statistischer Evidenz zwischen ähnlichen Wörtern sowie einen verbesserten Einbezug von textuellem Kontext – und damit verbunden flüssigeren Output, der dem menschlichen nie zuvor ähnlicher war. Das Geheimnis hinter solch neuronalen Systemen liegt gemäss Tiedemann (2017) im sog. *Representation Learning*, i.e. in der kontinuierlichen (nicht-symbolischen) Repräsentation von Wörtern, welche aus einem riesigen Datenset erlernt wird; nicht-symbolische Repräsentation meint, dass Wörter in neuronalen Systemen – im Unterschied zu ihren statistischen Vorgängern – unter Einbezug des Kontexts, in welchem ebendiese Wörter vorkommen, abstrahiert werden und dadurch als Konzepte (im virtuellen, mehrdimensionalen Raum) angeordnet werden können. Das diesem Ansatz zugrunde liegende Prinzip stammt aus der distributionellen Semantik und wurde von einem ihrer bedeutendsten Vertreter John R. Firth (1957) in folgendem Satz zusammengefasst: *You shall know a word by the company it keeps* – mit anderen Worten: Wörter, welche in ähnlichen Kontexten vorkommen, tendieren dazu, eine ähnliche Bedeutung zu haben; umgekehrt gilt, dass sich diejenigen Wörter, welche in (stark) unterschiedlichen Kontexten vorkommen, semantisch nicht gleichen.

Technisch gesehen werden dabei Kontextwörter zuerst in sog. *one-hot*-Vektoren enkodiert, was bedeutet, dass alle Wörter des Vokabulars der Grösse v in v-dimensionalen Vektoren abgebildet werden. Der Nachteil dieser *one-hot*-Vektoren besteht darin, dass jeweils nur eine Dimensionen der Vektoren relevante Informationen zum jeweiligen Wort enthält (i.e. die Zahl 1), während der grösste Teil der Dimensionen leer (i.e. 0) bleibt. Um diesen Engpass zu umgehen, wird das Vokabular des v-dimensionalen Raums in einen sog. d-dimensionalen Raum projiziert, sodass schlus-

sendlich diejenigen Wörter, welche in ähnlichen Kontexten vorkommen, ähnliche, verdichtete Vektoren haben und darum im virtuellen semantischen Raum näher beieinander zu stehen kommen.

2.1.2 Neuste Entwicklungen und Implementationen

Mit dem Release von Google's Neural Machine Translation System (GNMT)[1] im Jahre 2016 und nicht zuletzt seit der Markteinführung von DeepL[2] im August 2017 ist neuronale maschinelle Übersetzung nach dem anfänglichen Hype längst nicht mehr nur ein Buzzword, sondern hat sich als genuiner Fortschritt innerhalb der künstlichen Intelligenz erwiesen. Dabei sollte nicht vergessen werden, dass bereits in den 1980er- und 1990er-Jahren Forschung zu neuronalen Netzwerken und deren Anwendung in der maschinellen Übersetzung betrieben wurde (u.a. Waibel et al. (1991)) – mit dem Unterschied, dass damals keine für die Technologie genügend grossen Datensets verfügbar waren, um repräsentative und gleichzeitig qualitativ hochstehende Ergebnisse zu erzielen, und die Rechenkomplexität die damals verfügbaren Rechenressourcen schnell überstieg.

Die "moderne Auferstehung" habe, wie Koehn (2017) es schreibt, mit der Integration von neuronalen Sprachmodellen in traditionelle, statistische Übersetzungssysteme begonnen. Obwohl solche Ansätze wie dieser vielversprechend waren, wurden sie anfänglich nur bedingt angewendet, da (wieder) entweder die entsprechenden Rechenressourcen aus finanziellen Gründen nicht vorhanden waren (z.B. GPUs) oder es aber an Know-How, mit diesen Technologien umzugehen, fehlte.

Dass diese Engpässe mitunter vielleicht noch nicht ganz behoben, aber drastisch verbessert worden sind, dürfte klar sein – so mächtig und vielversprechend neuronale maschinelle Übersetzungssysteme heutzutage aber auch sein mögen, einige Herausforderungen und Performance-Probleme sind noch nicht gelöst. Koehn und Knowles (2017) geben in ihrem Paper *Six Challenges for Neural Machine Translation* einen diversifizierten und doch holistischen Überblick über die Performanz neuronaler Systeme verglichen mit den traditionellen statistischen Vorgängern im Hinblick auf sechs verschiedene Aspekte, von denen in den nächsten Zeilen diejenigen genauer erläutert werden, welche für diese Arbeit von Relevanz sind.

[1] `https://ai.google/research/pubs/pub45610`

[2] `https://www.deepl.com/press.html`

2.1.3 Aktuelle Herausforderungen

Wie bereits in der Einleitung erwähnt, steckt die Entwicklung von **Domänenadaptions-Methoden** für neuronale maschinelle Übersetzung noch in den Kinderschuhen. Koehn und Knowles kommen in ihren empirischen Untersuchungen zum Schluss, dass sich die *in-domain*-Performanz von statistischen und neuronalen Systemen nicht gross unterscheidet, die *out-of-domain*-Ergebnisse von neuronalen Systemen allerdings in fast allen Fällen (viel) schlechter sind als jene der statistischen; selbstverständlich ist diese Erkenntnis wertvoll, allerdings ist es m.E. etwas fraglich, Methoden zur Domänenadaption unter dem Aspekt der *out-of-domain*-Performanz zu evaluieren, da die Domänenadaption ja in erster Linie die Übersetzung von *in-domain*-Texten begünstigen sollte.

Die Autoren thematisieren als nächstes den Aspekt der **Trainingsdatenmenge**: Bei statistischen Modellen beobachten sie, dass der BLEU-Score bei der Verdopplung der Trainingsdaten (sowohl parallel als auch monolingual) in jedem Fall steigt; die Lernkurve neuronaler Systeme ist im Vergleich dazu bei Vergrösserung der Trainingsdaten zwar steiler, allerdings braucht diese Art von Systemen eine Mindestgrösse von ein paar Millionen Tokens, um überhaupt erst in Gang zu kommen und validen Output zu generieren – der Output neuronaler Systeme ist dabei (relativ) flüssig, allerdings entfernt er sich semantisch immer weiter vom Inputtext, je weniger Trainingsmaterial vorhanden ist. Für Domänenadaptions-Szenarien hat dies zur Folge, dass die Menge an Tuningmaterial genügend gross sein muss, um effektiv einen Einfluss auf die Qualität der Übersetzungen zu haben.

Ein weiterer Punkt, der gewissermassen mit der Trainingsdatenmenge zusammenhängt, bildet die **Übersetzung von seltenen Wörtern**, welche gemäss der Autoren in einem neuronalen Setting teilweise von der kleineren Vokabularsgrösse (als bei statistischen Systemen) abhängig ist. Interessanterweise schlagen neuronale Systeme die statistischen in der Übersetzung von sehr seltenen Wörtern, allerdings haben beide Systemarten Mühe mit infrequenten, stark flektierten Wörtern (wie z.b. Adjektive und Verben).

Des Weiteren ist seit den ersten *Encoder-Decoder*-Modellen gemeinhin bekannt, dass die **Übersetzung von langen Sätzen** Probleme bereitet. Um dem Abhilfe zu schaffen, wurden sog. *Attention*-Modelle eingesetzt, welche die Übersetzungsqualität von langen Sätzen verbesserte – allerdings nur bis zu einer bestimmten Satzlänge: Ab 60 Tokens sinkt die Qualität von neuronalem Output, während sie bei statistischen Modellen im Schnitt immer noch steigt. Interessant ist dabei, dass der neuronale Output stets relativ flüssig bleibt, allerdings oftmals zu kurze Übersetzungen gene-

riert werden, was den BLEU-Score natürlich erheblich nach unten zieht.

2.1.4 Nematus: a toolkit for neural machine translation

Die im Rahmen dieser Arbeit konzipierten neuronalen Modelle wurden anhand des Toolkits Nematus[3], welches hauptsächlich an der Universität Edinburgh entwickelt worden und frei verfügbar ist, trainiert. Dem Toolkit liegt eine sog. *Attentional Encoder-Decoder*-Architektur zugrunde, welche im Folgenden genauer erläutert wird:

Der Encoder besteht, wie es Sennrich et al. (2017) beschreiben, aus einem bidirektionalen rekurrenten neuronalen Netz (*Bidirectional Recurrent Neural Network (RNN)* mit sog. *Gated Recurrent Units (GRUs)*, welches eine Eingabesequenz $x = (x_1, ..., x_m)$ einliest und in einem nächsten Schritt je eine vorwärts ($\overrightarrow{h}_1, ..., \overrightarrow{h}_n$) und eine rückwärts ($\overleftarrow{h}_1, ..., \overleftarrow{h}_n$) gerichtete Sequenz von verborgenen Zuständen, *Hidden States*, berechnet. Für jedes Inputtoken werden sowohl der vorwärts ($\overrightarrow{h}_j$) als auch der rückwärts ($\overleftarrow{h}_j$) gerichtete verborgene Zustand konkateniert, woraus der Annotationsvektor h_j resultiert.

Der Decoder besteht wiederum aus einem rekurrenten neuronalen Netz, welches eine Ausgabesequenz $y = (y_1, ..., y_n)$ generiert, wobei jedes Token y_i aufgrund seiner rekurrenten verborgenen Zuständen, der zuvor vorhergesagten Tokens y_{i-1} und des Kontextvektors c_i vorhergesagt wird. Der Vorhersage jedes Ausgabetokens liegt also ein unterschiedlicher Kontextvektor zugrunde, welcher der gewichteten Summe aller Annotationen h_j entspricht. Während des Decodings des nächsten Tokens y_i wird dann das Gewicht – i.e. die Relevanz – jeder Annotation h_j durch ein sog. Alignmentmodell berechnet; das berechnete Gewicht entspricht dabei der Wahrscheinlichkeit, dass das Ausgabetoken y_i zum Eingabetoken x_j aligniert ist. Ausführlicher gesagt enthält also jede Annotation die Information über die gesamte Eingabesequenz, allerdings mit starkem Fokus auf diejenigen Teile dieser Sequenz, welche das i-te Inputtoken umgeben.

Attention, zu deutsch Aufmerksamkeit oder Beachtung, kann als eine Art Strategie betrachtet werden, anhand welcher versucht wird, die Performanz eines neuronalen Modells zu steigern. Diese Strategie gründet auf der Idee, dass das Voraussagen des nächsten Tokens in der Zielsprache auf dem letzten verborgenen Zustand – sprich dem "Speicher", der alle bisher kalkulierten Voraussagen beinhält – basiert.

Attention bietet dem neuronalen Modell also während des Decodings des nächsten

[3] `https://github.com/EdinburghNLP/nematus`

Wortes in der Zielsprache die Möglichkeit,"zurückzuschauen" und einem bestimmten verborgenen Zustand des enkodierten Eingabesatzes besondere Aufmerksamkeit zu schenken, i.e. auf die Information, welche in besagtem verborgenen Zustand enkodiert und für die Voraussage des nächsten Wortes relevant ist, zurückzugreifen; der Informationspfad ist damit für das Netz um einiges kürzer, als wenn es alle verborgenen Zustände durchlaufen (sowie zwischen relevanter und nicht-relevanter Information entscheiden) müsste. Letzteres Szenario ist vor allem bei langen Sätzen problematisch, da ohne den *Attention*-Mechanismus – wie es bei früheren neuronalen Modellen der Fall war – die gesamten Eingabeinformationen in einen sog. *fixed-dimensional* Vektor enkodiert werden, unabhängig davon, wie kurz oder lang die Eingabesätze tatsächlich sind. Es ist daher sowohl naheliegend als auch wissenschaftlich erwiesen, dass solch längenmässig fixe Vektoren bei langen Sätzen schnell sehr komplex werden und während des Decodings zu erhöhter Ungewissheit seitens der Entscheidungen – und damit schlussendlich zu qualitativ geringerem Output – des neuronalen Modells führen.

2.2 Domänenadaption

Wie im vorletzten Unterkapitel erwähnt, ist Domänenadaption bei neuronalen Systemen ein schwieriges Unterfangen. Sind die computergenerierten Übersetzungen zwar oft sehr fliessend und grammatikalisch (weitgehend) korrekt, so leidet, wie Khayrallah et al. (2017) es schreiben, die Adäquatheit der Übersetzung, i.e. die Reproduktion desselben semantischen Wertes, umso mehr. Nichtsdestotrotz haben sich im vergangenen Jahr einige Forschende vertieft mit dieser Thematik befasst und versucht, geschickte Lösungsansätze zu finden:

Luong und Manning (2015) entwickelten ihren ***Finetuning***-Ansatz in erster Linie für die maschinelle Übersetzung für Domänen gesprochener Sprache bzw. für ressourcenarme Sprachen. Nichtsdestotrotz ist ihre Arbeit aber in vielen anderen Forschungsprojekten und -ansätzen, die sich ausschliesslich mit der automatischen Übersetzung von geschriebenen Texten befassen, in Betracht gezogen und diskutiert worden. Die Technik, welche die beiden Autoren zur Domänenadaption verwenden, ist konzeptionell simpel und funktioniert folgendermassen: Das auf den *out-of-domain*-Trainingsdaten (4.5 Millionen Sätze) trainierte System wird, nachdem es mittels *Early Stopping* (u.a. anhand der Validierungsdaten) den qualitativ besten Punkt erreicht hat, für 12 Epochen nur noch auf den *in-domain*-Daten (200'000 Sätze) trainiert, wobei die *Learning Rate* nach jeweils zwei Epochen halbiert wird. Mit dieser simplen, aber effektiven Technik wird (zumindest in ihrem Szenario) eine

Verbesserung in der Qualität der übersetzten *in-domain*-Testdaten von 3.8 BLEU-Punkten erreicht.

Um das bei der *Finetuning*-Methode typischerweise schnell einsetzende *Overfitting* möglichst zu reduzieren, schlagen Miceli Barone et al. (2017) diverse sog. Regularisierungstechniken vor, welche zusätzlich die Qualität von *Transfer Learning* verbessern soll.

Ein weiteres Forscherpaar, das sich auf die Arbeit von Luong und Manning bezieht bzw. das *Finetuning*-Modell weiterentwickelt, sind Freitag und Al-Onaizan: Analog zu ersteren trainieren sie ein *Baseline*-System (bzw. verwenden sie ein "already trained baseline model"), führen dann mit nur einem kleinen Teil der *in-domain*-Daten das Training für wenige Epochen weiter, sodass sie ein sog. *Continue*-Modell erhalten, und kombinieren schlussendlich das *Baseline*- mit dem *Continue*-Modell zum ***Ensemble***-Modell, um möglichem *Overfitting* (des *Continue*-Modells) entgegenzuwirken. Interessanterweise verbessert sich durch diesen Ansatz nicht nur die Übersetzungsqualität der *in-domain*-Testdaten, sondern auch jene des *out-of-domain*-Testsets – wenn auch nur sehr minim (+0.6 bzw. +0.8 BLEU-Punkte).

Die Ergebnisse, welche in diesem Paper beschrieben und evaluiert werden, sind insofern von grosser Bedeutung, als dass sie nicht nur numerische Vergleichswerte (i.e. Verbesserung um BLEU-Punkte) sondern auch qualitative (Verbesserung gemäss "a human subjective evaluation metric") liefern; beide Vergleichswerte werden im Kapitel 6 zur Evaluation zu Rate gezogen.

Ein weiterer Ansatz zur Domänenadaption liefern Chu et al. (2017) mit sog. ***mixed Finetuning***. Die Autoren glauben dabei, dass *mixed Finetuning* das *Overfitting*-Problem zu umgehen vermag: Da in ihren Experimenten sowohl *Finetuning* als auch *mixed Finetuning* bereits nach 1 Epoche Domänen-adaptiven Trainings zu konvergieren beginnen, wurde das Training jeweils kurz nach 1 Epoche gestoppt. Während es bei der *Finetuning*-Methode nach 1 Epoche bereits zu gewissem *Overfitting* kam, schien dies beim *mixed Finetuning* nicht der Fall zu sein. Zudem nehme die Qualität der *out-of-domain*-Übersetzungen anhand des Domänen-adaptierten Systems bei ihrem Ansatz nicht ab, bei sowohl *Multi-Domain* als auch *Finetuning* sei dies allerdings der Fall. Die Autoren erläutern aber nicht nur Vorteile, sondern thematisieren auch einen spezifischen Engpass ihres Ansatzes, nämlich die längere Trainingsdauer bei der Domänenadaption – die Autoren begründen dies "as the time until convergence is essentially proportional to the size of the data used for fine tuning".

Eine neue Methode der Domänenadaption bei neuronalen Systemen, welche sich ***Cost Weighting*** nennt, stellen Chen et al. (2017) vor: Bei diesem Ansatz – wel-

chen die Autoren als "SMT domain adaptation technique" bezeichnen – werden ein Domänen-Classifier und ein Übersetzungsmodell simultan trainiert, wobei letzterer sowohl auf *in-domain*- als auch *out-of-domain*-Daten trainiert wird. Der Classifier soll dabei verwendet werden, um jedem Satz im gesamten Trainingsset eine Wahrscheinlichkeit, dass es sich um *in-domain*-Material handelt, zuweisen. Der bei jedem Satz berechnete *Cost* wird dabei gemäss der zuvor erwähnten Wahrscheinlichkeit gewichtet. Im Unterschied zu den drei vorherigen Ansätzen wird bei diesem hier keine eigentliche *Finetuning*-Strategie angewendet – vielmehr werden die *Cost*-Quotienten direkt gemäss der Ähnlichkeit eines jeden Satzes zum Validierungsset skaliert.

Dies, so die Autoren, beeinflusse das *Sequence-to-Sequence*-Modell zu Gunsten der *in-domain*-Daten und resultiere damit in einer verbesserten Performance bei der Übersetzung des *in-domain*-Testsets (ZH→EN: +1.2 BLEU-Punkte; EN→FR: + 0.8 BLEU-Punkte). Zudem sei so der Grundstein für mögliche *Onlineadaptation* gelegt.

Dass Domänenadaption in neuronalen Settings – wie bereits erwähnt – noch wenig erforscht ist, zeigt u.a. die Uneinigkeit, die bezüglich der Anzahl weiterer Epochen, während welchen die eigentliche Domänenadaption stattfindet, herrscht: Während Luong und Manning anhand 12 Epochen Domänen-adaptiven Trainings die besten Ergebnisse erhalten, erklären Sennrich und Kocmi (2016) in ihrem Handout zur Lab Session des Machine Translation Marathons 2016, man solle nur maximal 1 weitere Epoche auf den *in-domain*-Daten trainieren, da man ansonsten Gefahr laufe, dass das *Overfitting* schnell einsetzt. Zu diesem Schluss kommen indes sowohl Freitag und Al-Onaizan als auch Chen et al. und Chu et al. in den vorhin zitierten Papers.

Ein weiterer kritischer Punkt bildet die Definition von performanctechnischem Fortschritt: zu behaupten, dass ein gewisser Ansatz den besten Fortschritt erzielt, ist selbstverständlich legitim. Handelt es sich bei der Verbesserung aber nur um einen Sprung um nicht mal 1 BLEU-Punkt, stellt sich die Frage, inwieweit diese jeweiligen Erkenntnisse als Fortschritt (in Chen et al.'s (2017) Worten: "method obtaining the best improvement") bezeichnet werden sollten. Solch empirisch fundierten Erkenntnisse sind nämlich durchaus wertvoll, auch – oder gerade dann – wenn sie klar zeigen, welche Ansätze (unter den gegebenen Umständen) nicht oder nicht sehr gut funktionieren. Dass dennoch bei sehr vielen untersuchten Ansätzen in der Forschungsliteratur der performancetechnische Fortschritt wichtiger zu sein scheint, als die eigentliche(n) Erkenntnis(se), ist, meiner bescheidenen Meinung nach, schade. In dieser Arbeit soll darum der Fokus vermehrt auf qualitative Erkenntnisse – und nicht (nur) auf quantitative Forschritte – gelegt werden.

3 Datenaufbereitung

3.1 Domänenfremde Daten (out-of-domain)

Als domänenfremder Datensatz werden die Trainingsfiles vom *WMT17 NMT Shared Task: Machine Translation of News*[1] verwendet, die von den Organisatoren zur Verfügung gestellt werden; es handelt sich dabei um Ressourcen aus dem *Europarl Corpus*[2] sowie dem *UN Corpus*[3]. Der frei verfügbare WMT17-Datensatz besteht aus zwei Files, i.e. der deutschen und der englischen Version des Europarl Corpus mit je 1'920'209 Sätzen; die parallelen Files enthalten jeweils einen Satz pro Zeile.

3.1.1 Preprocessing

Als nächstes gilt es, den gesamten Datensatz zu tokenisieren und in *Byte Pairs* zu encoden, was in den nächsten Abschnitten erläutert wird. Sowohl die Tokenisierung (inkl. True Casing) als auch das *Byte Pair Encoding (BPE)* wurden anhand der Skripts und der Beispielkonfigurationen zur Vorverarbeitung durchgeführt, welche das Team um die *Edinburgh Neural MT Submission*[4] entwickelt und frei verfügbar bereitgestellt hat.

Der Konsistenz bzw. der Reduktion von potentiellen Fehlern halber sind zudem alle im deutschen Datensatz vorkommenden Eszett-Schriftzeichen *(ß)* durch die Sequenz *ss*, wie es nach Schweizer Rechtschreibung üblich ist, ersetzt worden. Diese Transliteration lässt sich insofern rechtfertigen, als dass es sich beim Eszett einzig um ein Graphem und kein eigenständiges, bedeutungsunterscheidendes Phonem der Deutschen Sprache handelt; zudem kommt im *in-domain*-Datensatz, der in Schweizer Hochdeutsch verfasst ist, auch kein Eszett vor.

[1] `http://www.statmt.org/wmt17/nmt-training-task/`

[2] `http://www.statmt.org/europarl/`

[3] `https://conferences.unite.un.org/uncorpus`

[4] `http://data.statmt.org/wmt17_systems/training/`

3.1.1.1 Tokenisierung

Nach erfolgreich durchgeführter Tokenisierung beträgt der deutsche Trainingssatz 48'760'058 Tokens (vorher: 44'675'235 Tokens), der englische 51'254'906 Tokens (vorher: 47'945'116 Tokens). Vor dem Truecasing, i.e. der Bestimmung sowie Wiederherstellung der inhärenten Gross- oder Kleinschreibung von Wörtern, sortiert das Skript alle leeren Zeilen und (zu) langen Sätze sowie Sätze mit einer hohen Source-Target-Ratio (i.e. einer hohen string-mässigen Übereinstimmung von Wörtern im Ausgangs- und Zielsatz) aus.

3.1.1.2 Truecasing

Der Truecaser wird nun auf der tokenisierten Korpusvariante trainiert, d.h. für alle Wörter wird berechnet, wie oft sie gross- oder kleingeschrieben, i.e. am Satz-/Zeilenanfang oder inmitten eines Satzes) vorkommen; diejenige Variante, die überwiegt, wird als die "wahre" (*true*) definiert und alle Vorkommnisse des jeweiligen Wortes werden in die wahre Ursprungsform umgewandelt. Nach erfolgtem Truecasing betragen die Trainingssätze 50'959'010 Tokens (DE) bzw. 53'575'330 Tokens (EN).

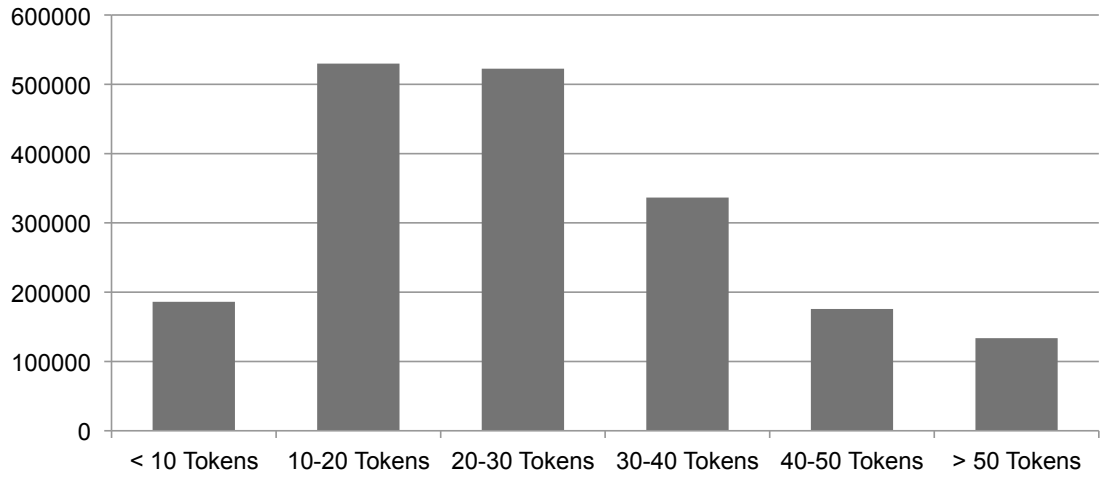

Abbildung 1: Anteil Satzlängen im *out-of-domain*-Trainingsset

3.1.1.3 Byte Pair Encoding

Neuronale maschinelle Übersetzungssysteme arbeiten typischerweise mit einem fixen, i.e. beschränkten Vokabular – Übersetzung ist aber eigentlich ein sog. *open vocabulary*-Problem, denn die Trainingsdaten, auf welchem ein maschinelles Übersetzungs-

system trainiert wird, werden – mögen sie auch noch so gross sein – nie alle möglichen Inputtokens enthalten. Um dieser Herausforderung entgegenzutreten, wurde in der Vergangenheit oft der Ansatz, unbekannte Wörter in einem Wörterbuch (*Dictionary*) nachzuschlagen, verfolgt; eine relativ simple und sicherlich effektive Methode, allerdings nur dann, wenn das unbekannte Wort sich auch tatsächlich im Wörterbuch befindet. Zudem ist nicht immer gegeben, dass sich Ausgangs- und Zielwort 1 zu 1 entsprechen – man denke dabei an deutsche (Mehrfach-)Komposita, deren Übersetzung oftmals aus mehr als einem Token besteht.

Eine andere, weitaus elegantere und zudem effizientere Möglichkeit, die in dieser Problematik Abhilfe verschaffen kann, ist das sog. *Byte Pair Encoding*: Eine simple, iterative Datenkompressionstechnik, in welcher die häufigsten Paare (Bigramme) konsekutiver Bytes mit einem neuen, ungebrauchten Byte dargestellt (i.e. *encoded*) werden. Der Algorithmus wurde erstmals 1994 von Philip Gage (1994) beschrieben.

Sennrich et al. (2016a) übernehmen bzw. adaptieren diesen Ansatz nun für die Sprachverarbeitung, genauer gesagt für die Wortsegmentation, wobei anstelle von Bytepaaren Zeichen bzw. Zeichensequenzen zusammengeführt werden. Ziel dieser Wortsegmentation ist es, unbekannte oder rare Wörter (im Trainingskorpus) in Subworteinheiten zu zerlegen, sodass die Übersetzungen ebendieser Wörter generiert werden können, obwohl die Wörter während des Trainings nicht im Datensatz enthalten waren. Die Intuition dahinter ist, wie Sennrich et al. in ihrem Paper beschreiben, dass die Übersetzung von gewissen Wörtern für einen professionellen Übersetzer insofern transparent sind, als dass er/sie diese Übersetzung intellektuell zu generieren vermag, ohne sie im Vorfeld zu kennen – basierend auf den dem Übersetzer bekannten, linguistischen Subworteinheiten wie Morphemen oder Phonemen. Als Wörter mit potentiell transparenter Übersetzung bezeichnen die Autoren u.a. Eigennamen (sofern Ausgangs- und Zielsprache über dasselbe Alphabet verfügen), Kognate und Lehnwörter sowie morphologisch komplexe Wörter (durch Affixation, Flexion oder Kompositabildung).

Dass aus der Umsetzung von *Byte Pair Encoding* zur Wortsegmentierung in der Praxis etwas andere, sprich weniger linguistische Subworteinheiten resultieren, zeigen die unten stehenden Beispiele, welche während des *Byte Pair Encodings* des WMT17-Trainingsdatensatzes generiert worden sind:

- Geburtshelfer: Ge@@ bur@@ t@@ shel@@ fer@@
- Teppichs: T@@ ep@@ p@@ ich@@ s
- Friedenskorps: Frieden@@ sk@@ or@@ ps

- Interimsetappe: Interim@@ se@@ ta@@ ppe
- lyrischen: l@@ y@@ r@@ ischen
- oberlehrerhaften: ober@@ lehr@@ er@@ haften
- kompromissfreie: kompromiss@@ freie
- Management@@ und Kontrollystem: Management@@ - und Kontrollsystem

Bei genauerer Inspektion wird ersichtlich, dass aus dem *Byte Pair Encoding* verschiedene Typen von Subworteinheiten resultieren können: Das Spektrum reicht von nicht-bedeutungstragenden bzw. nicht-morphologischen (linguistischen) Einheiten und Einzelbuchstaben (z.B. bur@@, shel@@, y@@) über morphemähnliche Zeichengruppierungen (z.B. Ge@@, er@@) bis zu fast eigenständigen Nomen (z.B. Frieden@@, kompromiss@@, Management@@).

3.1.2 Validierungs- und Testset

Eine grosse Herausforderung beim Trainieren von neuronalen maschinellen Übersetzungssystemen stellt das Stoppen des Trainings dar: Die Fehlerrate verkleinert sich während des Trainings kontinuierlich, allerdings setzt ab einem gewissen Punkt sog. *Overfitting* ein, was bedeutet, dass die Trainingsdaten vom System zu sehr memorisiert worden sind, sodass das System nicht mehr fähig ist, auf ungesehene Daten zu generalisieren. Um den idealen Zeitpunkt herauszufinden, zu welchem das Modell sowohl den bestmöglichen Fortschritt als auch seine grösstmögliche Generalisierungsfähigkeit erreicht hat, wird typischerweise ein sog. Validierungsdatenset verwendet. Dieses Datenset besteht typischerweise aus einem Anteil an *held-out*-Daten aus dem Trainingskorpus, i.e. Daten, welche dem Trainingsdatensatz zwar ähnlich, allerdings nicht für das Training verwendet worden (i.e. für das System neu) sind.

Für diese Arbeit wurde das beim *2007 Shared Task: Machine Translation for European Languages*[5] als Developmentset bereitgestellte Textfile verwendet, wobei die ersten 1000 der insgesamt 2000 Sätze als Validierungsset, die restlichen 1000 als Testset fungieren. Sowohl das Validierungsset als auch das Testset werden anhand der aus dem gesamten Trainingsset erlernten *Byte Pair Encodings* aufbereitet; beim Testset wurden allerdings nur die englischen Sätze anhand der *Byte Pair Encodings* segmentiert. Nach der erfolgten automatischen Übersetzung ins Deutsche werden dann in einem nächsten Schritt zuerst die Subworteinheiten wieder zu “echten” Wörtern

[5] `http://www.statmt.org/wmt07/shared-task.html`

zusammengefügt, dann folgen sowohl Detruecasing als auch Detokenisierung, sodass die automatische Übersetzung dann mit dem originalen deutschen Referenztext verglichen werden kann.

3.2 Domänenspezifische Daten (in-domain)

Als domänenspezifische Daten werden alle textuellen Daten aus Geschäftsberichten und Medienmitteilungen des Lebensversicherers *Swiss Life* bezeichnet, welche mit dem schriftlichen Einverständnis der Firma von der offiziellen Webseite gecrawlt worden sind. Es handelt sich dabei um parallele Texte, d.h. für jeden deutschen Bericht existiert das gekennzeichnete englische Pendant, was quasi den Schritt der Dokumentenalignierung – und damit eine potentielle Fehlerquelle – eliminiert. Die Geschäftsberichte bestehen jeweils aus etwa 200 Seiten und ergeben grob gerechnet etwa 1 Million parallele Tokens, bei den Medienmitteilungen, im Schnitt etwa 5 Seiten lang, sind es insgesamt rund 50'000 Tokens.

Inhaltlich geht es in den Geschäftsberichten v.a. um den Betriebsgewinn und die Massnahmen, die dazu geführt haben (relativ zum Vorjahr); zudem wird jeweils auf Aspekte der *Corporate Governance*, welche gemäss Bund[6] "Grundsätze und Regeln, mit deren Hilfe die Strukturen und das Verhalten der obersten Führungskräfte gesteuert und überwacht werden können" umfasst, eingegangen. Des Weiteren sind in den Geschäftsberichten jeweils die konsolidierten Jahresrechnungen enthalten, welche hauptsächlich aus Zahlen und Tabellen bestehen, allerdings regelmässig in längeren Fliesstextpassagen kommentiert werden.

Die Medienmitteilungen, welche alle Quartale publiziert werden, informieren über die (kurzfristige) Geschäftsentwicklung sowie über die an die Umsetzung der Unternehmensstrategie gekoppelten Massnahmen. Zudem wird jeweils über Geschäftsleitungs- oder Verwaltungsratsnominationen wie auch über Konferenzen für Investoren, Analysten und Medienschaffende informiert. Im Gegensatz zu den Geschäftsberichten handelt es sich bei den Medienmitteilungen fast ausschliesslich um Fliesstext.

Trotz der strukturellen Unterschiede können Geschäftsberichte und Medienmitteilungen als homogene Daten, sprich als zur selben Domäne zugehörig, angesehen werden, da sie sich inhaltlich relativ ähnlich sind.

[6] `https://www.kmu.admin.ch/kmu/de/home/praktisches-wissen/kmu-gruenden/firmengruendung/auswahl-rechtsform/aktiengesellschaft/was-bedeutet-corporate-governance.html`

3.2.1 Preprocessing

Die gesamte Aufbereitung der domänenspezifischen Daten umfasst die Konvertierung der PDF-Texte zu Rohtext (.txt); die anschliessende (erste) Bereinigungs- bzw. Korrekturphase; die Satzsegmentierung, i.e. ein Satz pro Zeile; Tokenisierung sowie Truecasing; Satzalignierung; *Byte Pair Encoding*. Alle Teilschritte werden im Folgenden detaillierter erläutert. Zudem soll auf potentielle Fehler aufmerksam gemacht werden, die sich aus der sequentiellen Abfolge der einzelnen Teilschritte ergeben könnten.

3.2.1.1 PDF to Text

Die Konvertierung der Geschäftsberichte und Medienmitteilungen (PDF) in Rohtext (.txt) erfolgte anhand des Kommandozeilentools `pdftotext`, welches im frei verfügbaren Toolkit `Xpdf`[7] enthalten ist. `pdftotext` liest ein beliebiges PDF und wandelt es in Rohtext um; dabei ist wichtig, dass die beiden Parameter `-raw` und `-enc 'UTF-8'` gesetzt sind, sodass sowohl die Umwandlung in "rohe" Zeilen als auch die richtige Textkodierung gewährleistet sind. Bilder und andere grafische Elemente werden eliminiert, was einerseits in Ordnung ist, da es sich nicht um textuellen Inhalt handelt; andererseits tauchen dadurch oft mitten im Fliesstext Bildunterschriften oder erklärende Worte zu Grafiken auf, welche relativ schwierig aufzuspüren sind, da sie bspw. nicht immer mit "Abbildung X:" beginnen. Zahlentabellen sind eine weitere Herausforderung, zumal deren Spalten- und Zeilenstruktur meist nicht wahrheitsgetreu in Rohtext übertragen wird. Allerdings ist dieser Verlust der Zahlentabellen im Rohtext nicht weiter tragisch, da das Preprocessingskript – wie in den oberen Abschnitten erwähnt – sowieso alle Zeilen mit einer hohen Source-Target-Ratio entfernt.

3.2.1.2 Satzsegmentierung

Da `Hunalign` als Input segmentierte sowie tokenisierte Sätze verlangt, wurden alle Texte mittels eines eigens kreierten Skripts segmentiert. Das Skript (siehe Appendix) ersetzt zunächst alle Zeilenumbrüche (`\n`) mit einem Leerzeichen (`\s`), dann werden alle Sätze des Texts mit der Funktion `sent_tokenizer`, welche im Package `ntlk.tokenize`[8] enthalten ist, segmentiert, sodass das Outputfile schlussendlich aus je einem Satz pro Zeile besteht. Ist die Satzsegmentierung einmal erfolgt, gilt

[7] `https://www.xpdfreader.com/about.html`

[8] `http://www.nltk.org/api/nltk.tokenize.html`

es auch hier, den Output genau zu überprüfen, um die Auswirkung von Segmentierungsfehler auf die weiteren Verarbeitungsschritte möglichst zu minimieren: Datumsangaben, bestehend aus *Tag, Punkt, Leerzeichen, Monat, Leerzeichen, Jahr* – z.B. *13. März 2002* – wurden häufig am Punkt voneinander getrennt, ebenso Ausdrücke wie *1. Quartal* oder aber Abkürzungen wie *bzw.* im Deutschen oder *e.g.* im Englischen. Um diese Fehlsegmentierung bzw. fehlgeschlagene (oder nicht erfolgte) Punktdesambiguierung rückgängig zu machen, wurde ein selbst geschriebenes Skript verwendet (siehe Appendix), welches die eben beschriebenen Fehler mittels regulärer Ausdrücke erkennt und korrigiert.

Eine weitere, zwar nicht direkt mit der Segmentierung in Verbindung stehende, aber dennoch nicht triviale Schwierigkeit / Fehlerquelle bilden v.a. in den deutschen Texten die Trennstriche (-), welche einerseits als "echte" Kompositabinder fungieren *(das Kollektivleben-Geschäft, die Swiss Life-Gruppe)* und darum so belassen werden müssen, andererseits in der PDF-Version die Funktion der Wort- bzw. Silbentrennung innehatten und bei der Konversion in Rohtext genau so übernommen worden sind *(des Investoren- tags)*. Letztere sind besonders problematisch, da nicht einfach alle Sequenzen von `-\s` entfernt werden können, weil ansonsten auch Fälle wie *Sach- und Haftpflichtversicherungsgeschäft* betroffen wären, indem sie zu *Sachund Haftpflichtversicherungsgeschäft* umgewandelt würden.

Die regulären Ausdrücke sind jeweils anhand des Onlinetools RegExr[9] getestet und verifiziert worden.

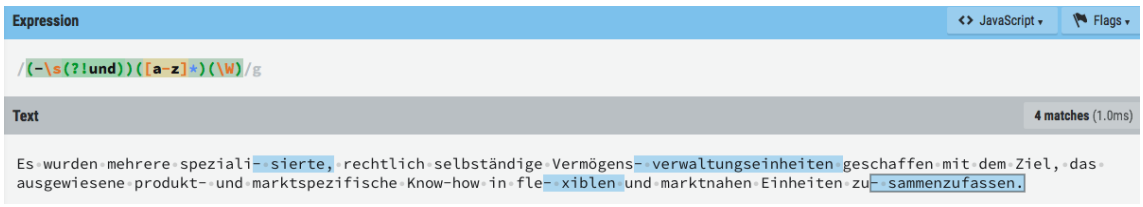

Abbildung 2: Testen der regulären Ausdrücke

3.2.1.3 Tokenisierung und Truecasing

Wie der *out-of-domain*-Datensatz wurden auch die domänenspezifischen Daten mit dem Preprocessingskript der Edingburgh-Truppe der Tokenisierung sowie dem Truecasing unterzogen. Die Datensätze betragen nach der Tokenisierung 834'293 (DE) bzw. 946'696 (EN) Tokens (vorher: 825'857 bzw. 951'538), nach dem Truecasing

[9] `https://regexr.com`

bestehen sie aus 946'926 (DE) bzw. 1'076'845 (EN) Tokens in je 38'585 Sätzen.

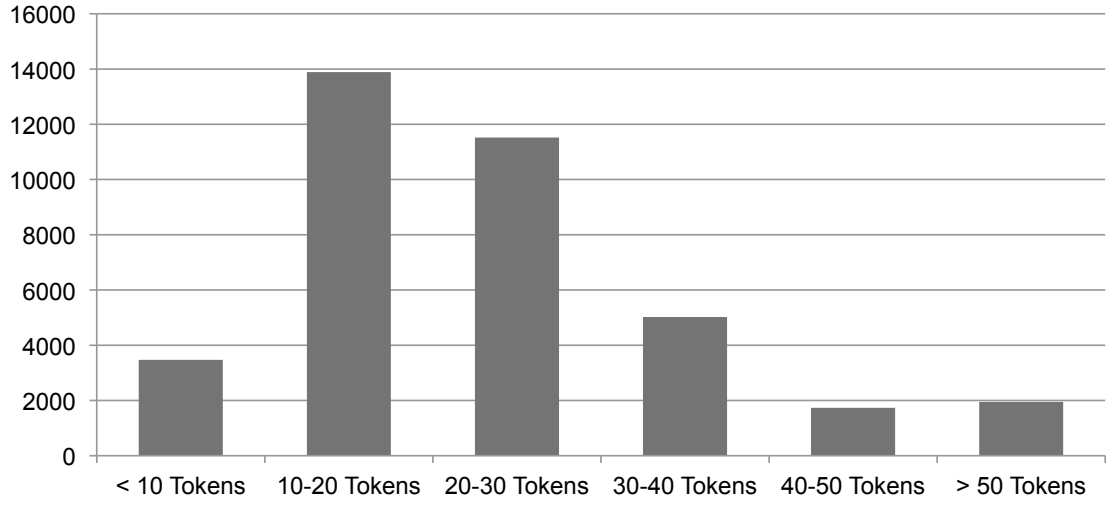

Abbildung 3: Anteil Satzlängen im *in-domain*-Trainingsset

3.2.1.4 Satzalignierung

Um mögliche (dokumentenübergreifende) Fehlalignierungen zu vermeiden, wurde die Alignierung mit `Hunalign`[10] mit allen korrespondierenden Dokumenten einzeln vorgenommen, d.h. die domänenspezifischen Einzeltexte wurden erst nach der Satzalignierung zu einem Korpus zusammengefügt. Nach jeder Satzalignierung (pro Dokument) wurden die Satz- bzw. Zeilenanzahl manuell auf Übereinstimmung überprüft und allfällige Korrekturen vorgenommen, sodass sich bei der Konkatenation aller Files (pro Sprache) keine Folgefehler ergaben, der bereinigte parallele Datensatz qualitativ also möglichst hochstehend ist.

3.2.1.5 Byte Pair Encoding (BPE)

Die Wortsegmentierung durch *Byte Pair Encoding* hat bei den domänenspezifischen Daten u.a. folgende Einheiten ergeben:

[10] `https://github.com/danielvarga/hunalign`

- Marktdurchdringung: Markt@@ durch@@ dr@@ ing@@ ung
- zukunftsbezogen: zukunfts@@ bezogen
- Bilanzergebnis: Bil@@ an@@ zer@@ geb@@ n@@ is
- Rentenanstalt: Renten@@ anst@@ alt
- Zinsänderungsrisiko: Zins@@ än@@ derungs@@ risiko
- CHF: C@@ H@@ F
- Berichtsperiode: Bericht@@ speriode
- Prämienverbilligung: Prämien@@ verb@@ illi@@ gung

Auch hier ist ersichtlich, dass das Spektrum an enkodierten Zeichen von bedeutungslosen, eher kurzen Zeichensequenzen (zer@@, än@@, illi@@) bis hin zu längeren, praktisch eigenständigen semantischen Einheiten (Renten@@, Bericht@@, Prämien@@) reicht.

3.2.2 Einbezug von domänenspezifischer Terminologie

Ein wichtiger qualitativer Aspekt bildet bei maschinellem Output die Konsistenz der (domänenspezifischen) Terminologie: Um erahnen zu können, wie gut – i.e. wie konsistent – spezifische, finanzwirtschaftliche Terme mittels eines domänenadaptierten Übersetzungssystem übersetzt werden, wurden die 100 häufigsten Terme als speziell domänenspezifisch definiert. Dazu wurde der gesamte deutsche domänenspezifische Datensatz mittels `TreeTagger`[11] mit Wortartentags versehen und anschliessend nach den 100 am häufigsten vorkommenden Adjektiv-Nomen-Kombinationen durchsucht.

Anschliessend wurde für diese 100 deutschen Mehrwortausdrücke die jeweils richtige englische Übersetzung festgelegt, wobei dazu die alignierten Sätze zu Rate gezogen worden sind. Schliesslich wurde für alle 100 deutschen Ausdrücke ein Abgleich mit dem NZZ-Korpus[12] gemacht, das als Referenz dient: Die gemeinhin als qualitativ hochstehende Zeitung charakterisierte NZZ besticht einerseits durch inhaltlich fundierte Artikel im Wirtschaftsteil, andererseits wird eine differenzierte und korrekte deutsche Sprache gepflegt; alle Adjektiv-Nomen-Ausdrücke, die sich nun in diesem Korpus wiederfanden, wurden dann von der Terminologieliste entfernt, sodass nur diejenigen Ausdrücke als domänenspezifische terminologische Einheiten

[11] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[12] `http://www.bubenhofer.com/korpusanalyse/MWE/Kombinationen/1995-2005/`

definiert werden, die tatsächlich nur im domänenspezifischen Datensatz der finanzwirtschaftlichen Texte aus der Versicherungsbranche vorkommen.

Die nachfolgende Tabelle enthält die 10 häufigsten Terme[13] (bestehend aus Adjektiv und Nomen), deren Frequenz (im domänenspezifischen Datensatz) sowie die englische Übersetzung:

#	Frequenz	Term DE	Term EN
1	638	finanzieller Vermögenswert	financial asset
2	375	konsolidierte Jahresrechnung	consolidated financial statement
3	361	ermessensabhängige Überschussbeteiligung	discretionary participation features
4	249	berufliche Vorsorge	occupational employee benefits
5	247	lokale Währung	local currency
6	202	weiteres Mandat	other appointment
7	191	aktivierte Abschlusskosten	deferred acquisition costs
8	191	aufgeschobene Vergütung	deferred compensation
9	183	finanzielle Verbindlichkeit	financial liability
10	314	eigene Aktie	own share

Tabelle 1: Ausschnitt Terminologie

Bei den meisten der 100 Terme handelt es sich nicht um Wörter, welche einen grossen Grad an Ambiguität aufweisen, weil sie in verschiedenen Domänen unterschiedliche Bedeutungen (und ggf. auch unterschiedliche Übersetzung) haben, wie es z.B. beim Wort *Bug* der Fall ist, welches in IT-spezifischen Texten als *Fehler*, in biologischen aber als *Käfer* übersetzt werden sollte. Vielmehr handelt es sich um Wörter bzw. Mehrwortausdrücke, welche (relativ) unambig sind, allerdings mehrere Übersetzungsvarianten aufweisen, die eigentlich alle korrekt sind: *Consolidated financial statement* wird von Google mit *konsolidierter Finanzbericht* übersetzt, während BingTranslator und DeepL die Variante *Konzernabschluss* wählen. Da Swiss Life wie viele Firmen auch allerdings eine konsistente Sprache pflegen – gerade bei finanztechnischen Reportings oder Richtlinien –, gilt das Credo, möglichst keine sprachliche Varianz in die Fachtermini hineinzubringen.

[13]Der gesamte Datensatz der 100 häufigsten Terme findet sich im Appendix.

3.2.3 Testset

Das domänenspezifische Testset, welches aus 1000 zufällig ausgewählten Sätzen (aus dem gesamten Swiss Life-Korpus) besteht, wurde ebenfalls anhand der aus dem *out-of-domain*-Trainingsdatensatz erlernten *Byte Pair Encodings* segmentiert.

Wie im vorherigen Unterkapitel erläutert wurde, spielt die terminologische Konsistenz bei der Übersetzung von firmenspezifischen, finanzwirtschaftlichen Texten eine wichtige Rolle und soll darum auch präzise evaluiert werden. Die untenstehende Tabelle enthält 10 der 100 als fachspezifische Termini definierten Mehrwortausdrücke, deren Frequenz im Testset sowie die deutsche Übersetzung. Die Evaluation der Übersetzung wird gerade bei denjenigen Begriffen interessant sein, welche im Testset mehrmals vorkommen – und darum im besten Fall auch jeweils gleich übersetzt werden.

#	Freq	Term EN	Term DE
1	13	operative costs	operative Kosten
2	11	comprehensive advice	umfassende Beratung
3	10	insurance reserves	versicherungstechnische Rückstellung
4	6	profitable growth	profitables Wachstum
5	5	Executive Officer	Executive Officer
6	4	deferred acquisition costs	aktivierte Abschlusskosten
7	4	statutory minimum distribution ratio	gesetzliche Mindestausschüttungsquote
8	1	consolidated balance sheet	konsolidierte Bilanz
9	1	forward-looking statement	zukunftsgerichtete Aussage
10	1	unknown risk	unbekannte Risiken

Tabelle 2: Ausschnitt Termini im Testset

Dasselbe gilt für Eigennamen, deren Übersetzungen ebenfalls möglichst konsistent sein sollten. Die untenstehende Tabelle zeigt die 10 am häufigsten auftretenden Eigennamen im Testset; auch hier wird es interessant sein, die jeweiligen Übersetzungen desselben Eigennamens auf ihre Konsistenz zu überprüfen.

#	Freq	Name
1	286	Swiss Life
2	40	La Suisse
3	39	Rentenanstalt
4	29	Banca del Gottardo
5	19	Rolf Dörig
6	17	AWD
7	13	SWX Swiss Exchange
8	12	Vaudoise
9	10	Helsana
10	3	Carsten Maschmeyer

Tabelle 3: Ausschnitt Eigennamen im Testset

Auf ein Validierungsdatenset ist indes bewusst verzichtet worden, da der ohnehin eher kleine Datensatz an domänenspezifischem Textmaterial nicht noch mehr verkleinert werden sollte; zudem fand die Domänenadaption nur während weniger, genau definierter Epochen statt, Validierung spielte demnach nur eine untergeordnete Rolle.

Die Repräsentativität der Testdaten ist insofern gewährleistet, als dass die Verteilung der verschiedenen Satzlängen sich sowohl im *out-of-domain*-Datensatz als auch in den *in-domain*-Trainings- und Testsets ungefähr gleich verhält, wie dem untenstehenden Diagramm entnommen werden kann.

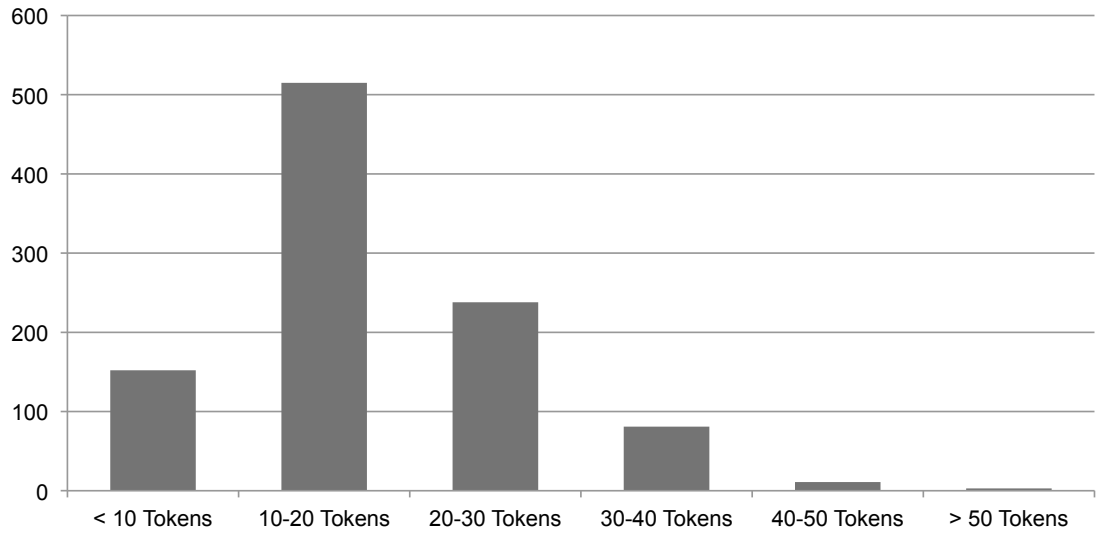

Abbildung 4: Anteil Satzlängen im Testset

3.3 Korpora-Statistiken

Baseline-Training	
Trainingsset	1920209
Validierungsset	1000
Testset	1000
Domänenadaption	
Trainingsset	37585
Testset	1000

Tabelle 4: Grösse der verwendeten Korpora (in Sätzen)

4 Baseline-System

4.1 Training

Als *Baseline*-System wurde ein Modell auf den *out-of-domain*-Trainingsdaten trainiert, welches in einem nächsten Schritt (siehe Kapitel 5) der Domänenadaption unterzogen wurde, wobei jeweils nach einer bestimmten Zahl an *Updates* anhand eines Validierungssets manuell validiert wurde. Ziel dieses testgetriebenen Vorgehens ist es, herauszufinden, inwiefern die Parameterkonfiguration(en) des *Baseline*-Systems sowie die Aufbereitung der Trainingsdaten (Preprocessing, *Byte Pair Encodings*) einen Einfluss auf die Performance des domänenadaptierten Systems hat.

4.1.1 Parameter

Für das Training des *Baseline*-Systems wurde grösstenteils die Parameterkonfiguration[1] von Sennrich et al. (2016b), welche damit Top-Resultate an der *First Conference on Machine Translation (WMT16)*[2] erzielten, übernommen; allfällige Abweichungen werden in den jeweiligen Unterkapiteln deklariert und genauer erläutert. Als Trainingsdaten wurde, wie in Kapitel 3 genauer beschrieben, der WMT17-Datensatz verwendet.

4.1.1.1 Datensets

Die drei Parameter, welche auf die während des Trainings gebrauchten Datensätze verweisen, sind `--datasets`, i.e. die parallelen tokenisierten, getruecasten und *Byte Pair*-enkodierten Trainingsfiles, `--dictionaries`, i.e. die Netzwerkvokabulare der Ein- und Ausgabesprache, sowie `--valid_datasets`, i.e. die ebenfalls parallelen tokenisierten, getruecasten und *Byte Pair*-enkodierten Validierungsfiles.

[1] `https://raw.githubusercontent.com/rsennrich/wmt16-scripts/master/sample/config.py`

[2] `http://www.statmt.org/wmt16/`

4.1.1.2 Word Embedding-Dimensionen und Hidden Layers

Analog Sennrich et al. wurden für das Training des *Baseline*-Systems *Word Embeddings* der Grösse 500 sowie *Hidden Layers* der Grösse 1024 verwendet.

Zum Vergleich: Freitag und Al-Onaizan verwenden für ihre Experimente 620 resp. 1000 "Zellen" pro RNN GRU layer, sowie ein Vokabular von je Wörtern 100'000; Chen et al. arbeiten mit *Word Embedding*-Dimensionen der Grösse 512 für die Eingabesprache und 1024 für die Ausgabesprache, sowie mit einem *Hidden Layer* der Grösse 1024. Die Vokabularsgrösse betrug im Setting ZH→EN 60'000, bei EN→FR 90'000.

4.1.1.3 Vokabularsgrösse der Ein- und Ausgabesprache

Während sich die Anzahl der *Word Embedding*-Dimensionen sowie der *Hidden Layers* in den vorher erwähnten Forschungsprojekten relativ bis sehr ähnlich sind, weichen sie in Punkto Grösse des Vokabulars der Ein- und Ausgabesprache deutlicher voneinander ab: Im Vergleich zu Freitag und Al-Onaizan sowie Chen et al. verwenden Luong und Manning ein eher kleines Vokabular der Grösse 50'000, dasjenige von Chu et al. beträgt sogar nur 32'000.

Nematus hält bei den beiden Parametern `--n_words_src` und `--n_words` als Standardwert `None`, was soviel heisst, dass das neuronale Netzwerk selbst die Grösse des Netzwerkvokabulars berechnet – im Falle der *out-of-domain*-Trainingsdaten, die in dieser Arbeit verwendet worden sind, beläuft sich die Grösse auf 19'000 Einträgen im Englischen und 35'000 im Deutschen. Im Preprocessingskript der Edinburgh-Truppe ist die Variable `bpe_operations` mit dem Kommentar[3] versehen, dass das Netzwerkvokabular leicht kleiner sein sollte als die Merging-Operationen, sofern die Operationen auf Basis der zusammengefügten Vokabulare (i.e. Ein- und Ausgabesprache) erlernt werden.

4.1.1.4 Back-Propagation und Gradient Descent Training

Dem Trainieren eines neuronalen Systems liegt der Grundsatz der iterativen Gewichtsanpassung zugrunde: Die gesamten Trainingsdaten werden, wie Koehn (2017) erklärt, wiederholt in das Netzwerk gespiesen; zudem wird die Übersetzung (eines Satzes), die zu einem gewissen Trainingszeitpunkt vom System generiert wird, mit

[3] Network vocabulary should be slightly larger (to include characters), or smaller if the operations are learned on the joint vocabulary [.]

dem Original verglichen. Als nächstes wird ein sog. *Error Term*, zu deutsch Fehlerrate, berechnet, welche gewissermassen die Abweichung des computergenerierten Satzes vom Original festhält. Diese Fehlerrate wird wiederum ins Netz, d.h. an jedes Neuron, zurückgespiesen, was im Fachjargon als *Back-Propagation* bezeichnet wird. Die Gewichte, anhand welcher die verschiedenen Neuronen miteinander verknüpft sind, werden in einem nächsten Schritt sukzessive angepasst: Die Formel zur Berechnung der aktualisierten Gewichte folgen dem Prinzip des *Gradient Descent Training*: Dabei wird der zurückgespiesene Fehlerwert eines jeden Neurons als Funktion der eingehenden Gewichte betrachtet. Um diesen Fehler zu minimieren, wird zuerst das mathematische Gefälle (*Gradient*) der Fehlerfunktion in Bezug auf jedes Gewicht berechnet. Dasjenige mathematische Gefälle, welches steiler ausfällt, ist richtungsweisender für die Anpassung der Gewichte.

4.1.1.5 Optimizer: Adam

Um den rohen Gradienten in ein Parameterupdate umwandeln zu können, wird der Netzwerkparameter `--optimizer=Adam` verwendet. Es handelt sich bei `Adam` um "a simple and computationally efficient algorithm for gradient-based optimization of stochastic objective functions", der von Kingma und Ba (2015) entwickelt worden ist. In Nematus ist `Adam` standardmässig als *Optimizer* definiert; Sennrich et al. verwenden in ihrer Samplekonfiguration allerdings den *Optimizer* `Adadelta`.

4.2 Validierung anhand Kreuzentropie und Early Stopping

Um während des Trainings die Entwicklung bzw. den optimalen Zustand des neuronalen Systems berechnen zu können, wird regelmässig anhand des in Kapitel 3 beschriebenen Validierungssets sowie einer sog. *Cost*-Funktion die Fehlerrate des Systems – gemessen anhand des Masses der Kreuzentropie – berechnet. Ziel ist es, das Training zum Zeitpunkt der tiefsten Fehlerrate zu stoppen, bevor allfälliges *Overfitting* einsetzt; die Fehlerrate fällt auf den verwendeten Trainingsdaten stetig, auf dem Validierungsset steigt sie aber nach einem gewissen Punkt wieder.

Wie bereits in Kapitel 2 erwähnt, wird das Training zum Zeitpunkt der tiefsten Fehlerrate nicht manuell, sondern typischerweise anhand von *Early Stopping* – i.e. dem automatischen Stoppen des Trainings, bevor es zu konvergieren beginnt – ausgeführt. Eine zentrale Rolle spielt hierbei der Parameter `--patience`, welcher in

der Keras-Dokumentation als “the number of epochs with no improvement after which training will be stopped”[4] bezeichnet wird; die Defaulteinstellung bei Nematus beträgt 10, d.h. das Training wird erst nach 10 Epochen, während denen kein (*Cost*-mässiger) Fortschritt stattfindet, gestoppt.

Untenstehende Grafik zeigt den Trainingsverlauf des *Baseline*-Systems, wobei der letzte Punkt der Kurve den Zeitpunkt, in welchem *Early Stopping* einsetzte, darstellt, was zwischen der 13. und 14. Epoche (dies entspricht ungefähr einer Trainingsdauer von drei Tagen) der Fall war. Die X-Achse zeigt die zeitliche Dimension (in *Updates*), die Y-Achse die Kreuzentropiewerte:

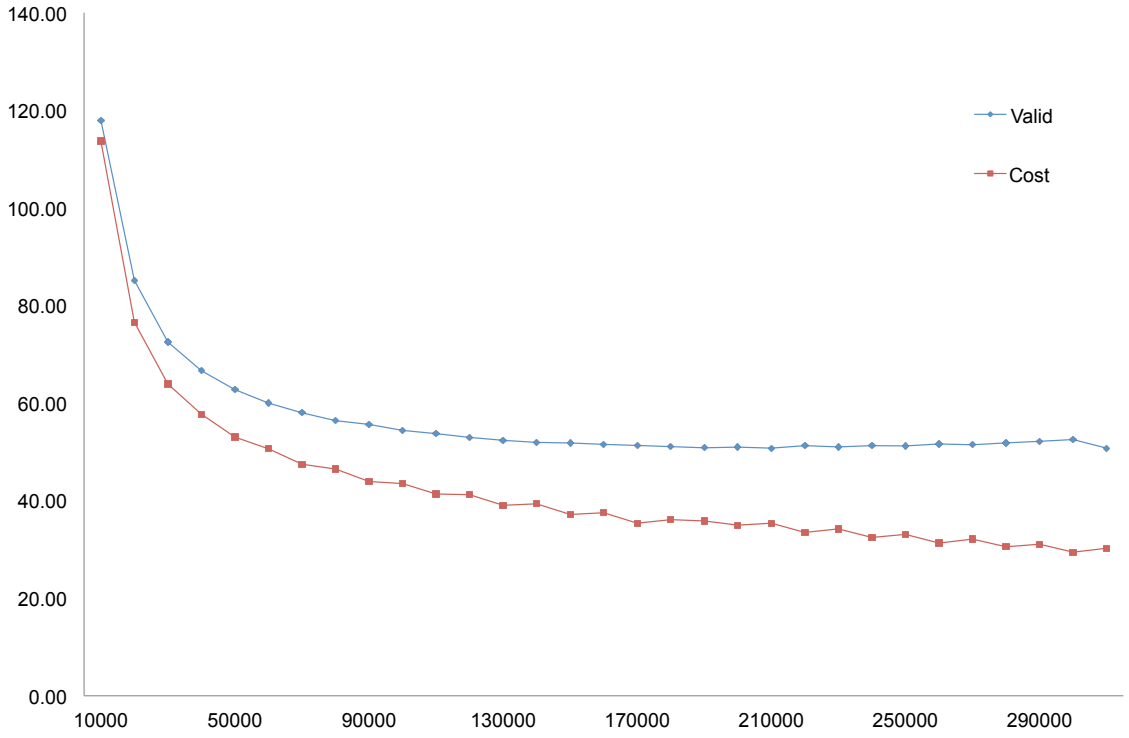

Abbildung 5: Trainingsverlauf

Ein ähnliches (wenn auch nur exemplarisches) Bild ist Koehn (2017) zu entnehmen:

[4] `https://keras.io/callbacks/#earlystopping`

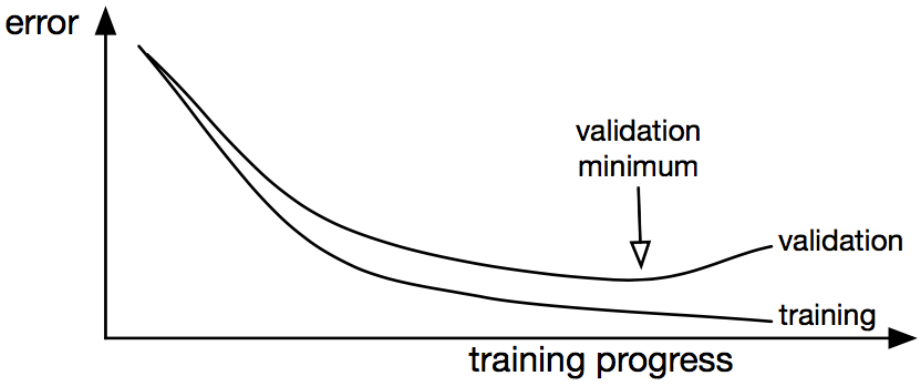

Abbildung 6: Exemplarischer Trainingsverlauf (gemäss Koehn (2017))

4.2.1 Shuffling, Sample- und Validierungs-Frequenz

Die folgenden Parameter `--shuffle_each_epoch=True`, `--sampleFreq=10000`, `--validFreq=10000` sowie `--saveFreq=30000` sind samt ihren Defaultwerten übernommen worden. Diese Parameter führen dazu, dass die Trainingsdaten nach jeder Epoche neu gemischt werden; dass jeweils nach 10000 Updates einige Samples (i.e. Ausgabesatz, Originalübersetzung sowie maschinelle Übersetzung, die das System zu besagtem Trainingszeitpunkt generiert) gezeigt werden; dass jeweils nach 10000 Updates automatisch validiert wird; sowie dass das Modell nach jeweils 30000 Updates gespeichert wird.

4.3 Evaluation

Nachfolgende Tabelle zeigt die Performance des für diese Arbeit trainierten *Baseline*-Systems im Vergleich zu den Ergebnissen, welche die in Kapitel 2 vorgestellten Autoren erzielten. Die Evaluation wurde einerseits auf dem Validierungsset (Dev), den *out-of-domain*-Testdaten (Test OOD) sowie den *in-domain*-Testdaten (Test ID) vollzogen, wobei alle Sets aus je 1000 Sätzen bestehen.

Nachfolgende Tabelle zeigt die Grösse der Trainingssets sowie die Übersetzungsqualität der *Baseline*-Systeme, welche von den in Kapitel 2 vorgestellten Autoren verwendet bzw. erzielt worden sind:

	Training OOD (M)	BLEU Test OOD	BLEU Test ID
Luong & Manning	4.5	–	25.6
Miceli Barone et al.	7.9	27.2	27.8
Freitag & Al-Onaizan	3.9	24.8	29.2
Chu et al.	1.0	37.1	2.6
Chen et al.	–	–	32.9
BL	1.9	22.0	14.0

Tabelle 5: Evaluation BL-Systeme

Mit den Defaultkonfigurationen von Nematus sowie 90'000 *Byte Pair-Merging*-Operationen und einer Vokabularsgrösse von 19'000 (EN) bzw. 35'000 (DE) schneidet das *Baseline*-Modell verglichen mit den Modellen der in Kapitel 2 vorgestellten Autoren verhältnismässig schlecht ab.

Im direkten Vergleich mit den Resultaten von Chu et al.'s *Finetuning*-Ansatz wird ersichtlich, dass der BLEU-Score bei der Übersetzung von *in-domain*-Testdaten in meinem *Baseline*-Modell um einiges höher ausfällt als bei Chu et al.'s *Baseline*-System, ihr Modell allerdings bei der Übersetzung der *out-of-domain*-Daten mein Modell klar übertrumpft. Dies ist insofern erstaunlich, als dass Chu et al. für das Trainieren ihres *Baseline*-Systems nur 1 Million parallele Sätze – also fast nur die Hälfte der für das Training dieses Modells eingesetzten Daten – verwenden; möglicher-weise kam es in ihrem Szenario wegen der kleinen Trainingsdatenmenge zu *Overfitting*, was den hohen *out-of-domain*- und den tiefen *in-domain*-BLEU-Score erklären könnte.

Die beste Übersetzungsqualität von *in-domain*-Testdaten erreichen Chen et al. – allerdings erwähnen sie in ihrem Paper, dass ihnen zu jedem Eingabesatz vier Referenzübersetzungen vorliegen, was die Wahrscheinlichkeit, dass die Tokens in der Referenz- und maschinellen Übersetzung übereinstimmen natürlich maximiert und der BLEU-Score darum potenziell höher ausfällt (wie in Kapitel 7 noch detaillierter diskutiert wird).

Die guten Resultate bei der Übersetzung der *out-of-domain*-Testdaten erstaunen allerdings bei Luong und Manning sowie Freitag und Al-Onaizan weniger, da beide eine weitaus grössere Trainingsdatenmenge verwenden.

5 Tuning (Domänenadaption)

5.1 Training

Um das domänenspezifische Modell zu trainieren, wurde der in Kapitel 2 beschriebene Ansatz des *Finetunings* verfolgt, bei welchem das bestehende *out-of-domain*-Modell für wenige Epochen auf dem domänenspezifischen Datensatz trainiert wird, wobei im Vergleich zu Miceli Barone et al. (2017) keine Massnahmen zur Reduktion des *Overfitting*-Risikos getroffen worden sind.

In den nachfolgenden Abschnitten wird genauer erläutert, welche Parameterkonfiguration für das Training des domänenspezifischen Modells verwendet worden ist.

5.1.1 Parameter

Die im vorherigen Kapitel erklärten Parameter sind für die Domänenadaption weitgehend übernommen worden – insbesondere *Dictionaries*, die Grösse des Netzwerkvokabuls sowie der *Word Embeddings* und *Hidden Layers* müssen mit denjenigen des originalen Modells übereinstimmen, wie dem Handout von Sennrich und Kocmi (2016) zur Lab Session des Machine Translation Marathon 2016 zu entnehmen ist.

Die *Byte Pair-Merging*-Operationen sind für die jeweiligen domänenspezifischen Trainingsdaten (gemäss Aufbereitung des *out-of-domain*-Trainingsmaterials) übernommen worden.

Die Parameter, welche auf die während des Trainings gebrauchten Datensätze verweisen, sind bei der Domänenadaption `--datasets`, i.e. die parallelen tokenisierten, getruecasten und bytepair-enkodierten *in-domain*-Trainingsfiles. Die Netzwerkvokabulare der Ein- und Ausgabesprache (`--dictionaries`) bleiben dieselben wie im domänenfremden Modell.

5.1.2 Training per Epochen

Da in der Forschungsliteratur zwar generell der Tenor gilt, Domänenadaption nur während weniger bzw. einer einzigen Epoche durchzuführen, allerdings dennoch – wie bereits in Kapitel 2 diskutiert wurde – keine absolute Einigkeit bezüglich der geeigneten Epochenanzahl herrscht, wurde das für diese Arbeit trainierte *Baseline*-Modell während 1, 2 und 5 zusätzlicher Epochen mit dem Domänen-spezifischen Datensatz trainiert; die *learning rate* wurde dabei konstant gelassen, im Gegensatz zum Ansatz von Luong und Manning, die sie nach jeweils zwei Epochen halbieren.

5.2 Evaluation

Die untenstehende Tabelle zeigt die Performance der domänenadaptierten Modelle sowohl auf dem domänenfremden als auch auf dem domänenspezifischen Testset. Wie auch bei den Ansätzen anderer Forscher zu sehen war, degradiert die Übersetzungsqualität des *out-of-domain*-Testsets kontinuierlich (hier um ca. 2 BLEU-Punkte), während diejenige des *in-domain*-Testsets schneller (hier quasi exponentiell), steigt. Die unterschiedlichen Grössen der Netzwerkvokabulare (Modell 1: automatisch, Modell 2: je 85000) scheinen keinen direkten Einfluss auf die Ergebnisse zu haben.

	Training ID (K)	BLEU Test OOD	BLEU Test ID
Luong & Manning	200	–	31.4
Miceli Barone et al.	206	30.8	31.5
Freitag & Al-Onaizan	194		
Epoche 1		25.4	32.2
Epoche 2		22.7	33.6
Epoche 20		20.3	30.5
Chu et al.	210	37.0	18.1
Chen et al.	–	–	34.1
Modell	37		
Epoche 1		21.20	**31.09**
Epoche 2		20.68	**33.93**
Epoche 5		18.94	**37.65**

Tabelle 6: Evaluation Domänen-adaptierte Systeme

Verglichen mit der Qualität der *in-domain*-Übersetzungen anhand des *out-of-domain*-

Baseline-Systems (14 BLEU-Punkte) hat sich jene der domänenadaptierten Systeme massiv verbessert; knapp 37'000 parallele – notabene qualitativ hochstehende – Sätze scheinen demnach auszureichen, um nach nur einer Epoche Domänen-adaptierten Trainings die Qualität der übersetzten domänenspezifischen Daten bedeutend zu verbessern. Zudem verbesserten sich durch die während des domänenadaptiven Trainings verwendeten Daten auch die Qualität der übersetzten *out-of-domain*-Sätzen. Im Gegensatz dazu muss aber auch angenommen werden, dass bei beiden Modellen nach mehr als einer zusätzlichen Epoche Domänen-adaptierten Trainings wahrscheinlich sehr schnell sog. *Overfitting* einsetzte.

Ziel der Evaluation im nächsten Kapitel ist es, zu ergründen, wie sich ein quantitativer Fortschritt von mehr als 10 bzw. 2 bzw. 4 BLEU-Punkten nun effektiv – speziell hinsichtlich der in Kapitel 3 definierten domänenspezifischen Termini sowie Eigennamen – qualitativ niederschlägt.

6 Qualitative Analyse

Innerhalb der maschinellen Übersetzungs-Community hat sich der BLEU-Score, bei dessen Berechnung für jeden Satz im maschinell generierten Output N-Gramme gebildet und diese dann mit den N-Grammen der menschlichen Referenzübersetzung(en) verglichen werden, über die letzten Jahre praktisch als De-facto-Standard für Evaluation etabliert. Es handelt sich bei BLEU durchaus um ein valides und praktisches Evaluationsmass, nicht zuletzt wegen seines schnellen und – zumindest mathematisch – nachvollziehbaren Feedbacks zur Qualität des maschinellen Outputs in einem Experimentiersetting. Kritische Stimmen wurden allerdings bereits wenig später nach der Publikation des Papers *BLEU: a method for automatic evaluation of machine translation* von Papineni et al. (2002) laut: So bemängelten bspw. Callison-Burch et al. (2006), dass ein höherer BLEU-Score nicht zwingend eine genuine Verbesserung in der Übersetzungsqualität darstellte, und plädierten darum u.a. dafür, BLEU nur spezifisch zur Nachverfolgung von inkrementellen Systemverbesserungen oder aber zum Vergleichen von verschiedenen, allerdings auf derselben Technologie (z.B. statistisch) basierenden Systemen zu verwenden.

Je näher die maschinelle Übersetzung zur menschlichen Referenz, desto höher die Qualität des maschinellen Outputs – dies ist die grundlegende Idee, worauf der BLEU-Score, i.e. der Vergleich von N-Grammen des maschinellen Outputs mit denjenigen des menschlichen Referenztextes, basiert. Dieser Ansatz steht allerdings in direktem Widerspruch zur Natur der Sprache, denn für einen Satz existieren unzählige, wenn nicht unendlich viele Übersetzungen – anders gesagt kann der semantische Wert eines Satzes in einer Übersetzung absolut adäquat wiedergegeben werden, auch wenn die maschinelle Übersetzung im Vergleich zur menschlichen Referenz eine gewisse Variation in den einzelnen Textbausteinen, i.e. Wörtern, aufweist. Nüchtern betrachtet berücksichtigt BLEU zwar ein gewisses Mass an Variation, allerdings nur, sofern mehrere Referenzübersetzungen gegeben sind, die besagte Variation quasi belegen können. BLEU straft aber bspw. eine Übersetzung, welche semantisch absolut korrekt ist, deren Wörter aber nur wenig identisch mit der Referenzübersetzung sind, ab; umgekehrt kann es sein, dass BLEU einen (scheinbar) kleinen Unterschied im maschinellen Output nur wenig straft, obwohl der Satz durch

den kleinen Unterschied eine komplett andere Bedeutung erhält.

Mit dem Aufkommen von neuronalen Übersetzungssystemen – welche zunehmend Übersetzungen auf Basis von Subwörtern (*Byte Pair Encodings*) oder Zeichen generieren – wird der BLEU-Score zwangsläufig (erneut) auf die Probe gestellt werden. Die Forschergemeinde sollte sich daher m.E. nicht nur dem Ziel, BLEU-Werte weiter zu steigern, widmen, sondern sich zwingend damit auseinandersetzen, welche anderen Masse zur Evaluation von neuronalen Systemen und deren Output geeignet wären, um BLEU entweder zu ergänzen oder gegebenenfalls vollständig abzulösen. Es kann – wie Callison-Burch et al. bereits 2006 bemängelten – nicht sein, dass sich Forschungserfolge allein auf quantitative Ergebnisse abstützen, qualitative Aspekte aber vollkommen vernachlässigt werden.

Klubička et al. (2017) halten richtigerweise fest, dass automatische Evaluationsmetriken nicht anzuzeigen vermögen, ob ein gewisses sprach- – oder gar domänen- – spezifisches Phänomen vom Übersetzungssystem adäquat übersetzt worden ist. Die Autoren machen deshalb Gebrauch vom *Multidimensional Quality Metrics (MQM) Framework*[1], einer von Lommel et al. (2014) entwickelten Evaluationsmethode, anhand welcher taskspezifische, hierarchische Fehlerkategorien zuerst definiert und dann in einem nächsten Schritt auf den maschinell übersetzten Text angewendet werden. Diese Evaluationsart erlaubt es, einen gegebenen Text nicht nur auf verschiedene Fehlerarten zu prüfen, sondern innerhalb dieser Fehler zwischen kleinen, grossen und sehr kritischen Fehlern zu unterscheiden, und damit eine viel präzisere Auswertung zu erhalten.

Selbstverständlich sind solch manuelle Auswertungen sehr zeitintensiv und kostspielig und werden demnach in der Praxis nur eher selten durchgeführt. Um dem Anspruch, eine fundierte Evaluation durchzuführen, gerecht werden zu können, werden in dieser Arbeit nichtsdestotrotz einige *in-domain*-Testsätze zusätzlich zur automatischen BLEU-Evaluation manuell geprüft. Die (sprach-/domänenspezifischen) Phänomene, die es zu untersuchen gilt, werden im nächsten Unterkapitel genauer erläutert.

6.1 Qualitätsparameter

Die grundlegenden Fragen, die es vor der qualitativen Evaluation zu klären gilt, lauten: Wie ist Qualität in einem Domänenadaptions-Szenario mit firmenspezifischen Texten zu definieren? Welche Aspekte sind – zusätzlich zum indikativen BLEU-Score

[1] `http://www.qt21.eu/mqm-definition/definition-2015-06-16.html`

– zentral, um ein objektives Urteil bzgl. der Qualität des maschinellen Outputs fällen zu können? Sind alle Aspekte gleichwertig zu behandeln oder gibt es welche, die bei der Qualitätseinschätzung wichtiger, sprich mehraussagend sind und darum mehr ins Gewicht fallen sollten?

Koby et al. (2014) definieren Qualität bei (maschinellen) Übersetzungen folgendermassen: "A quality translation demonstrates **accuracy** and **fluency** required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs."

Die fettgedruckten Begriffe sind dabei folgendermassen zu verstehen:

- Accuracy (Genauigkeit): Wie gut repräsentiert der Text in der Ausgabesprache den Informationsgehalt des Eingabetextes?
- Fluency (Sprachgewandtheit des Systems): Erfüllt der Text in der Ausgabesprache die Merkmale Grammatikalität, Klarheit und Format?

Ferner halten Lommel et al. fest, dass die Übersetzungsqualität nur danach beurteilt werden kann, ob eine Übersetzung den spezifischen Anforderungen und dem kommunikativen Zweck entspricht. Dass die Anforderungen an Genauigkeit und Sprachgewandtheit innerhalb der verschiedenen End-User variieren können, hat zwangsläufig zur Folge, dass auch qualitative Auswertungen nie vollständig sind – umso wichtiger scheint es daher, maschinellen Output aus der sowohl quantitativen als auch qualitativen Perspektive zu beurteilen, um einer allumfassenden, objektiven und validen Evaluation möglichst nahe zu kommen.

Aus diesen Überlegungen und dem *MQM Issue Types*-Katalog, der im Paper von Lommel et al. enthalten ist, leiten sich folgende qualitativen Evaluationsdimensionen ab, die in dieser Arbeit berücksichtigt werden, wobei die einzelnen Dimensionen der Einfachheit halber nicht unterschiedlich gewichtet werden:

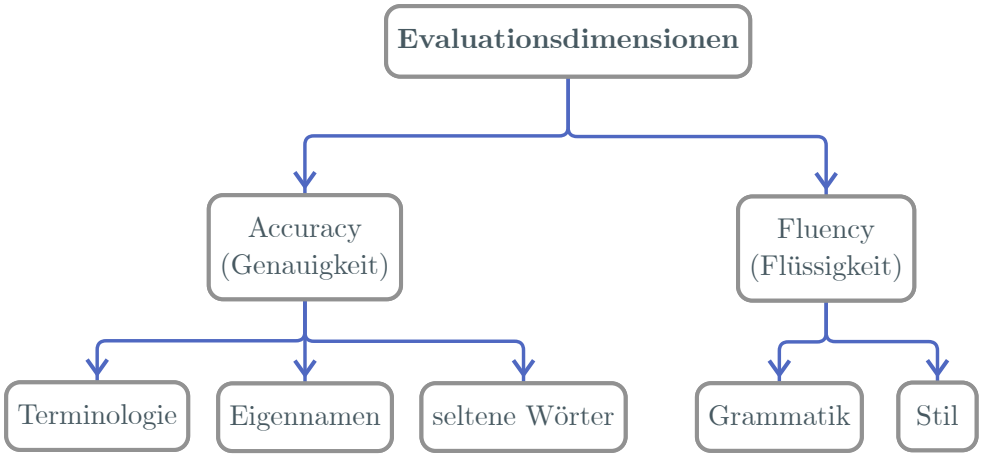

Abbildung 7: Dimensionen der qualitativen Analyse

Der Klarheit halber wird in der Evaluation folgendermassen auf die drei domänenadaptierten Modelle verwiesen:

- **DA1** → Modell mit 1 Epoche domänenadaptiven Trainings
- **DA2** → Modell mit 2 Epochen domänenadaptiven Trainings
- **DA5** → Modell mit 5 Epochen domänenadaptiven Trainings

6.1.1 Accuracy (Genauigkeit)

Es scheint mir an dieser Stelle wichtig, klarzustellen, dass das Beurteilen der Beispielssätze hinsichtlicher ihrer semantischen Nähe zur jeweiligen Referenz ein schwieriges Unterfangen ist, welches nur mit einem gewissen Grad an Fachwissen – über welches ich (zum heutigen Zeitpunkt) nicht verfüge – adäquat bewältigt werden kann. Deshalb wird der Aspekt der Genauigkeit nur in der Terminologie, bei Eigennamen sowie bei seltenen Wörtern (i.e. isolierten Einheiten) überprüft.

6.1.1.1 Terminologische Konsistenz

Die Übersetzungsqualität der zehn als domänenspezifische Termini definierten Kombinationen aus Adjektiv und Nomen ist, wie der untenstehenden Tabelle zu entnehmen ist, eher dürftig. Ferner scheint sich auch kein klarer Zusammenhang zwischen

der Frequenz der Ausdrücke im Trainingsset und des F1-Wertes herauszubilden. Im Schnitt verbessert sich die Qualität der Übersetzungen von DA1 zu DA2 zwar, allerdings fällt sie dann bei DA5 wieder minim.

#	Terminus	Freq Train	Freq Test	F1 DA1	F1 DA2	F1 DA5
1	operative Kosten	44	13	71%	75%	67%
2	umfassende Beratung	31	11	92%	95%	90%
3	versicherungstechnische Rückstellung	57	10	18%	75%	89%
4	profitables Wachstum	33	6	83%	73%	83%
5	Executive Officer	198	5	80%	80%	80%
6	aktivierte Abschlusskosten	140	4	100%	100%	75%
7	gesetzliche Mindest-ausschüttungsquote	101	4	–	40%	40%
8	konsolidierte Bilanz	100	1	–	100%	100%
9	zukunftsgerichtete Aussage	61	1	25%	67%	–
10	unbekannte Risiken	46	1	33%	100%	100%
	Durchschnitt			**10%**	**18%**	**17%**

Tabelle 7: Ausschnitt Termini im Testset

Die z.T. sehr tiefen Werte resultieren natürlich auch daraus, dass die Termini mit anderen Worten umschrieben (bzw. übersetzt) worden sind: So wurden z.B. *umfassende Beratung* öfters mit *umfassende Lösung* oder gar *umfassende Einzelversicherungslösung*, *operative Kosten* mit *Betriebskosten* übersetzt, was in allen Fällen semantisch sicherlich korrekt ist, allerdings nicht dem Ziel, formal konsistente Übersetzungen zu generieren, entspricht.

6.1.1.2 Konsistenz bei Eigennamen

Während die Mehrheit der Eigennamen relativ konstant gut übersetzt werden, bildet der Name *Rentenanstalt* eine Ausnahme, denn obwohl er im Trainingsset im Vergleich eher häufig erscheint, fehlt er oft bei den Übersetzungen und zieht damit aufgrund der vielen *False Negatives* den F1-Score nach unten.

Bei genauerer Betrachtung derjenigen Zeilen im übersetzten Testset, welche den Na-

men (gemäss Referenztext) eigentlich beinhalten sollten, wird allerdings ersichtlich, dass DA1 *Rentenanstalt* des öfteren mit *Rentenstalt* oder *Rentenstint* übersetzt; dies nimmt allerdings mit jeder weiteren Epoche ab, bis DA5 nur noch die korrekte Variante *Rentenanstalt* generiert. Hinzu kommt, dass bei allen *in-domain*-Daten, die vor 2004 erschienen sind, der Name *Swiss Life* nicht alleine, sondern oft in der kombinierten Variante *Swiss Life / Rentenanstalt* vorkommt. Die Übersetzungen dieses Doppelnamens lauten bei DA1 oft *Swiss Life / Swiss Life* oder eben *Swiss Life /* in Kombination mit einer der falschen *Rentenanstalt*-Versionen.

#	Eigennamen	Freq Trainingsset	Freq Testset	F1 DA1	F1 DA2	F1 DA5
1	Swiss Life	8398	286	97%	96%	96%
2	La Suisse	51	40	92%	97%	96%
3	Rentenanstalt	263	39	14%	46%	69%
4	Banca del Gottardo	214	29	98%	97%	98%
5	Rolf Dörig	261	19	97%	100%	100%
6	AWD	552	17	97%	94%	97%
7	SWX Swiss Exchange	34	13	100%	100%	100%
8	Vaudoise	19	12	87%	96%	92%
9	Helsana	6	10	95%	95%	95%
10	Carsten Maschmeyer	78	3	80%	80%	80%
	Durchschnitt			**86%**	**90%**	**92%**

Tabelle 8: Ausschnitt Eigennamen im Testset

Die obigen Resultate zeigen, dass die Übersetzung von Eigennamen mit weiteren Epochen domänenadaptiven Trainings im Schnitt besser werden, wobei der Fortschritt zwischen DA2 und DA5 nicht mehr so hoch ausfällt, wie es zwischen DA1 und DA2 der Fall war. Es lässt sich allerdings keine klare Tendenz zwischen der Frequenz der Eigennamen im Trainingsset und der Entwicklung des F1-Wertes ableiten.

6.1.1.3 Übersetzung von seltenen Wörtern

Wie Sennrich et al. (2016a) richtigerweise festhalten, tendieren seltene Wörter oft dazu, wichtige Informationsträger im Satz zu sein – sprich der eigentliche semantische Wert eines Satzes ist nur dann sichergestellt, wenn (mindestens) die zentrale

Information adäquat übersetzt wird.

Nachfolgende Tabelle zeigt 10 zufällig ausgewählte Wörter, welche sowohl im Trainings- als auch im Testset selten (i.e. weniger als 15 Mal) vorkommen, und die Übersetzungen, welche DA1, DA2 und DA5 daraus generieren:

EN	share issue	average price over three months
DE	Aktienemission	Drei-Monats-Durchschnittspreis
DA1	Aktie-Aktie	**Durchschnittspreis von mindestens drei Monaten**
DA2	Anstieg	Durchschnittspreis von mindestens drei Monaten
DA5	(nicht übersetzt)	über drei Monate eine Prämienzahlung
EN	(within) the EU	portfolio of business activities
DE	(im) EU-Raum	Geschäftsportefeuille
DA1	(nicht übersetzt)	**Portfolio des Geschäfts**
DA2	(nicht übersetzt)	Portfolio des Geschäfts
DA5	(nicht übersetzt)	Portfolio an die Geschäftsaktivitäten
EN	liability lines	commission payments
DE	Haftpflichtversicherungsgeschäft	**Kommissionszahlungen**
DA1	Verbindlichkeiten	Kommissionszahlungen
DA2	Absicherung	Kommissionszahlungen
DA5	(nicht übersetzt)	Kommissionszahlungen
EN	revaluation reserves	annuity conversion rate
DE	Neubewertungsreserven	Rentenumwandlungssatz
DA1	Rückstellungen	Rentenwandlungen
DA2	Rückversicherungsreserven	Rentenaufwandlungssatz
DA5	Bewertungsreserven	Rentenwandlungssatz
EN	supplementary	substantially
DE	überobligatorisch	substanziell
DA1	**ergänzend**	**besonders stark**
DA2	zusätzlich	besonders stark
DA5	zusätzlich	besonders stark

Tabelle 9: Ausschnitt seltener Wörter im Testset

Auf den ersten Blick scheint der Informationswert bei etwa der Hälfte der Übersetzungen (fett gedruckt) erhalten geblieben zu sein (von der Geläufigkeit dieser generierten Begriffe einmal abzusehen), auch wenn praktisch keine der Übersetzungen 1 zu 1 der Referenz entspricht. Die genauere Betrachtung der Kontexte, in welchen die obigen Wörter (im Trainingsset) vorkommen, zeigt folgendes:

Das Wort *überobligatorisch* bspw., ein Begriff aus der Rechtssprache, dessen englisches Pendant *supplementary* ist und vom Modell mit *ergänzend* übersetzt wird, kommt praktisch nur in Zusammenhang mit dem Wort *BVG* vor (*überobligatorisches BVG-Geschäft*, *überobligatorischer Teil der zweiten Säule*), die Begriffe bilden also eine Kollokation. *Ergänzend* mag eine absolut korrekte Übersetzung des Begriffs *supplementary* sein, allerdings ist eine Äusserung wie *ergänzendes BVG-Geschäft* nicht geläufig und damit – zumindest in dieser Domäne – formal eigentlich falsch.

Besonders interessant ist, dass die Übersetzung des Wortes *Kommissionszahlungen*, welches im gesamten Trainingskorpus nur 1 Mal vorkommt, 1 zu 1 mit der Referenz übereinstimmt – allerdings handelt es sich hier um ein deutsches Kompositum, das im Englischen aus denselben Wörtern (und sogar derselben Wortreihenfolge) besteht, und damit ein idealer Kandidat für *Bytepair Encoding* ist.

Wie der Tabelle zu entnehmen ist, scheint sich die Übersetzungsqualität von seltenen Wörtern – im Unterschied zu den Eigennamen – mit zunehmenden Epochen domänenadaptiven Trainings nicht zu verbessern.

6.1.2 Fluency (Sprachgewandtheit)

Um einen Eindruck zu haben, wie sich die Sprachgewandtheit der Sätze mit zunehmender Epochenanzahl entwickelt, sind insgesamt 10 Sätze unterschiedlicher Länge aus dem Testset zufällig ausgewählt worden. Diese Auswahl ist dabei nicht als (anzahlmässig) repräsentative Stichprobe zu verstehen, sondern dient lediglich dazu, einen konkreteren Einblick in die qualitative Performanz des Systems zu gewähren.

6.1.2.1 Grammatikalische Korrektheit

Kurze Segmente (weniger als 10 Tokens)
In den Segmenten mit weniger als 10 Tokens finden sich nicht nur abgeschlossene Sätze, sondern auch viele Titel (z.B. eines neuen Absatzes in einer Medienmitteilung). Bei genauerer Betrachtung dieser Titel wird ersichtlich, dass das Modell in denjenigen Fällen, in welchen im zu übersetzenden englischen Titel ein Verb (oft im Präteritum, i.e. Simple Past) vorkommt, Übersetzungen generiert, welche ebenfalls eine Verbform enthalten; in einigen Fällen wird sogar am Ende der Tokensequenz ein Punkt generiert, da das Modell dies – basierend auf dem bereits berechneten satzähnlichen Segment – vermutlich als das nächst logische Token berechnet. Entspricht bspw. Beispielssatz 1 formal nicht 100%ig der Referenz (bzw. dem englischen Eingabesatz), so ist er grammatikalisch gesehen absolut intakt.

EN	Allocation to reserves for policyholder bonuses doubled to CHF 1.7 billion
DE	Zuweisung an Überschussreserven auf CHF 1,7 Milliarden verdoppelt
DA1	Die Zuweisung für Rückstellungen für Versicherungsnehmer hat sich auf CHF 1,7 Milliarden verdoppelt.
DA2	Die Zuweisung von Rückstellungen für Versicherungsnehmer hat sich auf CHF 1,7 Milliarden verdoppelt
DA5	Die Zuweisung an Rückstellungen für Überschussanteile der Versicherungsnehmer konnte auf CHF 1,7 Milliarden verdoppelt werden.
EN	Embedded value up 20% to CHF 10.7 billion
DE	Embedded Value um 20% auf CHF 10,7 Milliarden erhöht
DA1	Der Embedded Value von 20% auf CHF 10,7 Milliarden
DA2	Der Embedded Value von 20% auf CHF 10,7 Milliarden
DA5	Embedded Value um 20% auf CHF 10,7 Milliarden

Tabelle 10: Übersetzung von Titeln

Diejenigen Segmente, bei welchen es sich um komplette Sätze handelt, werden zumeist bereits ab der ersten Epoche gute Ergebnisse generiert, wie in den untenstehenden drei Beispielen ersichtlich ist; spätestens ab der fünften Epoche nimmt die Qualität in Punkto grammatikalische Korrektheit allerdings wieder (mehr oder minder stark) ab.

EN	Swiss Life is renowned for reliability and quality.
DE	Swiss Life steht für Zuverlässigkeit und Qualität.
DA1	Swiss Life ist bekannt für Zuverlässigkeit und Qualität.
DA2	Swiss Life ist für Zuverlässigkeit und Qualität bekannt.
DA5	Für Swiss Life ist Swiss Life bekannt für Zuverlässigkeit und Qualität.
EN	Further details will be provided towards the end of April.
DE	Weitere Informationen zur Finanzierung folgen gegen Ende April.
DA1	Weitere Angaben werden bis Ende April vorliegen.
DA2	Weitere Angaben werden bis Ende April vorliegen.
DA5	Weitere Angaben dazu finden sich bis Ende April.
EN	The relevant exposure respresents a mere 0.1% of investments.
DE	Der Anteil entsprechender Anlagen liegt bei lediglich 0.1%.
DA1	Das betreffende Exposure stellt lediglich 0.1% der Kapitalanlagen dar.
DA2	Das betreffende Exposure ist nur 0.1% der Kapitalanlagen.
DA5	Das betreffende Exposure entspricht lediglich 1-1% der Investitionen.

Tabelle 11: Kurze Segmente

Mittlere Segmente (zwischen 20 und 40 Tokens)

Die drei Beispielssätze zeigen, dass DA1 und DA2 die grammatikalisch besten ("flüssigsten") Übersetzungen zu generieren vermögen – dabei sind insbesondere die Phrasen *per Ende des Monats März* (DA2) sowie *Währungseffekte infolge des stärkeren Euros* (DA1) beeindruckend, welche durch den Einsatz des Genitivs stilistisch fast eleganter erscheinen als die jeweilige Referenz. Dagegen sind die Übersetzungen von DA5 (in den Beispielen 1 und 2) nicht nur zunehmend ungrammatisch, sondern fallen (in 1 und 3) auch um einiges kürzer aus als der englische Eingabesatz sowie die von DA1 und DA2 generierten Sequenzen.

EN	Banca del Gottardo turned in a good performance , which was reflected in its segment result of CHF 168 million.
DE	Die Banca del Gottardo erzielte ein gutes Ergebnis, was sich in einem Segmentergebnis von CHF 168 Millionen niederschlug.
DA1	Die Banca del Gottardo hat sich in einer guten Leistung gewandelt, die sich in ihrem Segmentergebnis von CHF 168 Millionen niedergeschlagen hat.
DA2	Die Banca del Gottardo hat sich in einer guten Leistung gewandelt, was in ihrem Segmentergebnis von CHF 168 Millionen berücksichtigt wurde.
DA5	Die Banca del Gottardo erzielte eine gute Performance, die in ihrem Segmentergebnis von CHF 168 Millionen berücksichtigt wurde.
EN	The equity exposure of 6.2% at the end of March was lowered once again in April and stood at 2.8% at the end of June.
DE	Die Aktienquote von 6,2% per Ende März wurde im April wieder reduziert und belief sich per Ende Juni auf 2,8%.
DA1	Die Eigenkapitalrechnung von 6,2% Ende März wurde im April erneut abgesenkt und lag per Ende Juni bei 2,8%.
DA2	**Die Aktienquote von 6,2% per Ende des Monats März wurde im April wieder abgesenkt und lag per Ende Juni bei 2,8%.**
DA5	Die Eigenkapitalquote von 6,2% Ende März wurde im April wieder gesenkt und lag per Ende Juni bei 2,8% auf 2,8%.
EN	The units sold are fully consolidated up to the completion of the transaction concerned, but only their net contribution to the net profit is included in the consolidated statement of income.
DE	Die verkauften Bereiche werden bis zum Abschluss der jeweiligen Transaktion voll konsolidiert, aber in der konsolidierten Erfolgsrechnung nur noch mit ihrem Nettobeitrag zum Reingewinn berücksichtigt.
DA1	**Die verkauften Einheiten werden bis zum Abschluss der betreffenden Transaktion voll konsolidiert, aber nur ihr Nettobeitrag zum Reingewinn ist in der konsolidierten Erfolgsrechnung enthalten.**
DA2	Die verkauften Einheiten werden bis zum Abschluss der betreffenden Transaktion voll konsolidiert, aber nur ihr Nettobeitrag zum Reingewinn ist in der konsolidierten Erfolgsrechnung enthalten.
DA5	Die verkauften Einheiten werden bis zum Abschluss der betreffenden Transaktion voll konsolidiert, aber in der konsolidierten Erfolgsrechnung nur noch mit ihrem Nettobeitrag zum Reingewinn erfasst.

Tabelle 12: Übersetzung von Segmenten zwischen 20 und 40 Tokens

Lange Segmente (mehr als 50 Tokens)

Gemäss Koehn und Knowles (2017) sinkt die Übersetzungsqualität bei sehr langen Sätzen – allerdings ist sie bis zu einer Satzlänge von 60 Tokens immer noch höher als bei traditionellen statistischen Systemen. Zudem würden ab einer gewissen Länge zu kurze Übersetzungen generiert. Im hiesigen Testset sind nur 3 englische Eingabesätze enthalten, welche aus mehr als 50 (i.e. 59 bzw. 81 bzw. 73) Tokens bestehen, deren maschinelle deutsche Übersetzungen allerdings nicht viel signifikant kürzer ausfallen, wie der untenstehenden Tabelle zu entnehmen ist:

EN	However, at the request of the FOPI, it will be agreed, that if, nevertheless, the bank is sold to a third party within the next three years, 50% of any realised gain above the agreed transfer price would be **remitted to** Swiss Life/Rentenanstalt, thereby benefitting the policyholders.
DE	Auf Begehren des BPV wird dennoch vereinbart, dass – falls die Bank wider Erwarten innert der nächsten drei Jahre an eine Drittpartei verkauft würde – 50% des realisierten Gewinns aus einem über dem für den Transfer fixierten Preis erzielten Verkaufserlös der Versicherungsgesellschaft Rentenanstalt/Swiss Life zufliessen und damit auch den Versicherten zugute kommen würde.
DA1	Auf Antrag der **FOPI** wird jedoch vereinbart, dass die Bank in den nächsten drei Jahren an eine Drittpartei verkauft wird, 50% der realisierten Gewinne über den vereinbarten Transferkurs **auf** Swiss Life bzw. **auf** die Versicherungsnehmer gewinnen.
DA2	Auf Antrag der FOPI wird aber auch vereinbart, dass, wenn die Bank in den nächsten drei Jahren **zu** einer Drittpartei verkauft wird, 50% der realisierten Gewinne über den vereinbarten Deckungspreis auf die Swiss Life-Swiss Life übertragen werden, was den Versicherungsnehmern zugute kommt.
DA5	Auf Antrag des FOPI wird jedoch vereinbart, dass, wenn die Bank in den nächsten drei Jahren **zu** einem Dritten verkauft wird, 50% der realisierten Gewinne über den vereinbarten Transaktionskurs an die Rentenanstalt / Swiss Life abrufbar sind.

Tabelle 13: Übersetzung von Segmenten mit mehr als 50 Tokens

Wie an den Beispielen zu sehen ist, generieren DA1 und DA5 tendenziell zwar eher kürzere Übersetzungen, wobei DA2 dem englischen Eingabe- bzw. dem deutschen Referenzsatz am nächsten zu sein scheint. Auffällig sind in diesen Beispielen die fehlerhaft übersetzten deutschen Präpositionen (fett gedruckt). Zudem bereitet das Akronym *FOPI* Schwierigkeiten, was allerdings nicht erstaunt, da es im Trainingskorpus nur fünf Mal vorkommt.

Ein allumfassendes Urteil bzgl. der besten Qualität zu fällen, scheint bei dieser Auswahl schwierig, wobei der beste Satz vermutlich von DA2 generiert wurde. Fakt ist, dass die Übersetzungen bei zunehmender Epochenzahl grammatikalisch gesehen nicht besser ausfallen.

6.1.2.2 Stil

Ein in der deutschen (Schrift-)Sprache häufig verwendetes Wortzeichen bildet der sog. Ergänzungsstrich, mittels welchem bei zwei oder mehreren aufeinanderfolgenden Komposita, die über dasselbe Grundwort verfügen, das Grundwort ersetzt wird – wie z.B. im Satz *Wir betreuen Privatkunden und Firmenkunden.* , der mit Ergänzungsstrich zu *Wir betreuen Privat- und Firmenkunden.* – und damit um einiges leserlicher – wird.

Auch im domänenspezifischen Trainingskorpus finden sich viele solcher verkürzten

Kompositionen wieder, welche bzgl. ihres Informationsgehaltes eigentlich auch als domänenspezifische Termini definiert werden könnten; eine Auswahl davon, welche auch im Testset vorhanden ist, zeigt nachfolgende Tabelle.

#	Kompositum	Freq Training
1	Versicherungs- und Kapitalanlageverträge	64
2	Privat- und Firmenkunden	57
3	Finanz- und Risikomanagement	16
4	Leben- und Vorsorgegeschäft	14
5	Sonder- und Währungseffekte	11
6	Finanzierungs- und Holdinggesellschaften	4
7	Vermögensverwaltungs- und Bankgeschäft	2
8	Emissions- und Kotierungsprospekt	0
9	Krankentaggeld- und Unfallversicherung	1
10	Motorfahrzeug-, Sach- und Haftpflichtversicherungsgeschäft	0

Tabelle 14: Ausschnitt verkürzte Komposita im Testset

Mehr als die Hälfte dieser Komposita kommt mehr als einmal im Trainingskorpus vor, woraus vermutet werden kann, dass es sich um Kollokationen handelt, welche möglichst identisch auch in den maschinell übersetzten Sätzen vorkommen sollten, sodass die stilistische Konsistenz der spezifischen Textsorte gewahrt wird.

Die Evaluation der ausgewählten Sätze zeigt, dass auch hier kein Fortschritt zwischen DA2 und DA5 auszumachen ist: Die Übersetzungen des zweiten Beispielssatzes, welcher das Kompositum *Versicherungs- und Kapitalanlageverträge* beinhält, sind bereits ab DA1 korrekt, vermutlich aufgrund der hohen Freqzenz des Kompositums im Trainingskorpus. Ebenso verhält es sich mit dem Kompositum *Finanz- und Risikomanagement*, welches im Trainingskorpus eine vergleichsweise mittlere Frequenz aufweist. Die Komposition *Finanzierungs- und Holdinggesellschaften*, welche sehr selten im Trainingsset vorkommt, wird bereits ab DA1 zumindest verständlich übersetzt, erst DA2 vermag die exakt selbe Buchstabenfolge der Referenz zu generieren; ab DA5 wird nicht nur die Übersetzung der Komposita, sondern auch der gesamte Satz an sich schlechter.

EN	Thomas Müller is an outstanding leader with great experience in financial and risk management.
DE	Thomas Müller ist eine überzeugende Führungspersönlichkeit mit grosser Erfahrung im Finanz- und Risikomanagement.
DA1	Thomas Müller ist ein herausragender Führer mit grossen Erfahrungen im **Finanz- und Risikomanagement**.
DA2	Thomas Müller ist ein herausragender Führer mit grossen Erfahrungen im Finanz- und Risikomanagement.
DA5	Thomas Müller ist eine ausgezeichnete Führungsführerin mit Erfahrung im Finanz- und Risiko- Management.
EN	Policy fees received under insurance and investment contracts went up 37% to CHF 298 million.
DE	Die Policengebühren aus Versicherungs- und Kapitalanlageverträgen erhöhten sich insgesamt um 37% auf CHF 298 Millionen.
DA1	Die im Rahmen von **Versicherungs- und Kapitalanlageverträgen** erhaltenen Politien beliefen sich auf insgesamt 37% auf CHF 298 Millionen.
DA2	Die in Versicherungs- und Kapitalanlageverträgen erhaltenen Policen beliefen sich auf insgesamt 37% auf CHF 298 Millionen.
DA5	Unter Versicherungs- und Kapitalanlageverträge entfielen Swiss Life-Verträgen um 37% auf CHF 298 Millionen.
EN	The result from the “Other” segment (mainly financing and holding companies) together with “Eliminations” totalled CHF 4 million (previous year: CHF -73 million).
DE	Für das Segment Übrige, das vor allem Finanzierungs- und Holdinggesellschaften umfasst, und die Eliminationen ergab sich insgesamt ein Beitrag von CHF 4 Millionen (Vorjahr: CHF -73 Millionen).
DA1	Das Ergebnis aus dem Segment “Übrige” (hauptsächlich die Finanz- und Holdinggesellschaften) zusammen mit “Eliminationen” CHF 4 Millionen (Vorjahr: CHF -73 Millionen).
DA2	Das Ergebnis aus dem Segment “Übrige” (hauptsächlich **Finanzierungs- und Holdinggesellschaften**) zusammen mit “Eliminationen” CHF 4 Millionen (Vorjahr: CHF -73 Millionen).
DA5	Das Ergebnis des Segments “Übrige” (vor allem die Finanzierung und Holding-Gesellschaften) zusammen mit “Energie” belief sich auf CHF 4 Millionen (Vorjahr: CHF -73 Millionen).

Tabelle 15: Ausschnitt verkürzte Komposita

6.2 Synthese

Aus der obigen Analyse geht hervor, dass sich der quantitative Sprung von 31.09 (DA1) auf 33.93 (DA2) BLEU-Punkte qualitativ in Punkto Genauigkeit, Sprachgewandtheit und Stil der Sätze deutlich spürbarer äussert, als von 33.93 (DA2) auf 37.65 (DA5); die Qualität der Ergebnisse stagniert zwischen DA2 und DA5 nicht

nur, sondern sie wird z.T. sogar schlechter.

Diese Erkenntnis ist von grosser Wichtigkeit, denn sie zeigt auf der einen Seite – sozusagen auf der Metaebene –, dass mathematische und linguistische Analysen von Sprache durchaus zu unterschiedlichen Ergebnissen führen können (was zu beweisen war), auf der anderen Seite kann so genauer eruiert werden, welche Phänomene denn überhaupt charakteristisch für die spezifische Domäne (in der jeweiligen Sprache) sind und wie gut die Übersetzung ebendieser funktioniert.

7 Fazit

In dieser Masterarbeit wurde anhand des Toolkits Nematus ein neuronales *State-of-the-Art*-Übersetzungssystem auf generischen Daten trainiert, sowie zu einem späteren Zeitpunkt mittels simpler *Finetuning*-Methodik an die finanzwirtschaftliche Domäne, welche in der Literatur bisher nur wenig Erwähnung fand, adaptiert. Eine differenzierte Analyse sollte schlussendlich Aufschluss über die Performanz des domänenspezifischen Systems geben und konkrete Ergebnisse zeigen. Der Fokus der gesamten Arbeit lag einerseits auf der Aufbereitung des domänenspezifischen Korpus', andererseits auf der qualitativen Evaluation.

Zur Orientierung dienten dabei die drei in der Einleitung definierten Forschungsfragen:

1. Wie gut funktioniert Domänenadaption (anhand eines eigens kreierten Korpus' à 37'000 parallelen Sätzen) für finanzwirtschaftliche Texte?
2. Durch welche (linguistischen) Charakteristika zeichnet sich die finanzwirtschaftliche Domäne aus?
3. Wie konsistent werden diese in den maschinellen deutschen Übersetzungen beibehalten?

Wie in den vorherigen Kapiteln anhand quantitativer und qualitativer Auswertungen gezeigt worden ist, funktioniert Domänenadaption mit rund 37'000 *in-domain*- und nur 1.9 Millionen parallelen *out-of-domain*-Sätzen auf das Testset à 1'000 Sätzen gut; es ist allerdings anzunehmen, dass aufgrund der eher kleinen Datensets bereits nach 1 - 2 Epochen zusätzlichen, domänenadaptiven Trainings *Overfitting* einsetzt. Quantitativ ist zwischen den drei domänenadaptierten Modellen DA1, DA2 und DA5 ein BLEU-mässiger Fortschritt von 2 bzw. 4 Punkten auszumachen, was in der Literatur eigentlich als substantieller Unterschied gilt.

Die qualitative Analyse von einigen zufällig ausgewählten Sätzen unterschiedlicher Länge hat allerdings ergeben, dass die beiden Modelle DA1 und DA2 bzgl. der konsistenten Übersetzung von domänenspezifischen Fachtermini, Eigennamen und stilistischen Mitteln sowie bzgl. der Genauigkeit und Flüssigkeit der Sätze bessere Ergeb-

nisse erzielten als Modell DA5. Selbstverständlich stehen die wenigen manuell ausgewerteten Sätze in keinem Verhältnis zum gesamthaft verwendeten Testset und sind in gewisser Hinsicht auch nicht vollständig objektiv, aber sie ermöglichen zumindest das genaue Beobachten von allfälligen Tendenzen und sprach- bzw. domänenspezifischen linguistischen Charakteristika.

Nebst diesen spezifischen Eigenschaften können anhand der qualitativen Analyse aber auch Unregelmässigkeiten im Trainingsmaterial festgestellt werden: Der Ausdruck *Swiss Life/Rentenanstalt* bspw. kommt in dieser Form nur in den deutschen Sätzen vor, während die Entsprechung in den englischen Sätzen *Rentenanstalt/Swiss Life* lautet. Eine andere Variation lässt sich z.B. in der Segmentbezeichnung ausmachen, welche im Englischen *the segment "Other"* – mit Anführungs- und Schlusszeichen – lautet, im Deutschen aber ohne diese (*das Segment Übrige*) geschrieben wird. Dass solche Inkonsistenzen die Qualität eines Übersetzungssystems beeinflussen, je nachdem wie stark sie im Trainingsset vertreten sind, dürfte klar sein. Nichtsdestotrotz besteht die Gefahr, dass die gezielte und detaillierte Auseinandersetzung mit Texten angesichts der stark – und in naher Zukunft möglicherweise noch stärker – datengetriebenen Modellen vermehrt in den Hintergrund rückt.

Die Technologie, die hinter den maschinellen Übersetzungssystemen steckt, mag sich in den letzten Jahren durchaus verändert und zu einem signifikanten Qualitätsanstieg beigetragen haben. Fakt ist aber, dass dieser Qualitätsanstieg nur dann – sowohl in der Forschung als auch in der Industrie und nicht zuletzt in der Gesellschaft – nachhaltig gewährleistet werden kann, wenn Qualität nicht nur mit Zahlen, sondern ergänzend dazu anhand von Worten definiert, diskutiert und mit konkreten empirischen Beispielen belegt wird.

Quellen- und Literaturverzeichnis

Chris Callison-Burch, Miles Osborne und Philipp Koehn (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistic,* 249–256, Trento.

Boxing Chen, Colin Cherry, George Foster und Samuel Larkin (2017). Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation,* 40–46, Vancouver.

Chenhui Chu, Raj Dabre und Sadao Kurohashi (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers),* 385–391, Vancouver.

John R. Firth (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*, 1-32, Blackwell, Oxford.

Markus Freitag und Yaser Al-Onaizan (2016). Fast domain adaptation for neural machine translation, *CoRR.*

Philip Gage (1994). A new algorithm for data compression. *The C Users Journal,* 12(2): 23–38.

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post und Philipp Koehn (2017). Neural lattice search for domain adaptation in machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing,* Vol. 2: Short Papers, 20–25, Taipei, Taiwan.

Diederik P. Kingma und Jimmy Ba (2015). Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations,* San Diego, USA.

Filip Klubička, Antonio Toral Ruiz und M. Víctor Sánchez-Cartagena (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics,* 108(1):121–132.

Geoffrey Koby, Paul Fields, Daryl Hague, Arle Lommel und Alan Melby (2014). Defining translation quality. *Revista Tradumàtica: tecnologies de la traducció,* 413–420.

Philipp Koehn (2017). Statistical Machine Translation. Cambridge University Press, unpublizierte 2. Auflage.

Philipp Koehn und Rebecca Knowles (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation,* 28–39, Vancouver.

Arle Lommel, Aljoscha Burchardt und Hans Uszkoreit (2014). Multidimensional quality metrics: A flexible system for assessing translation quality. *Revista Tradumàtica: tecnologies de la traducció,* 455–463.

Minh-Thang Luong und Christopher D. Manning (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation,* 76–79, Da Nang, Vietnam.

Kishore Papineni, Salim Roukos, Todd Ward und Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics,* 311–318, Philadelphia, USA.

Rico Sennrich (2013). *Domain Adaptation for Translation Models in Statistical Machine Translation.* Doktorarbeit, Universität Zürich, Institut für Computerlinguistik, Zürich.

Rico Sennrich und Tom Kocmi (2016). Lab Session – Neural Machine Translation. Handout zum *Machine Translation Marathon 2016*, Prag.

Rico Sennrich, Barry Haddow und Alexandra Birch (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Vol. 1: Long Papers, 1715-1725, Berlin.

Rico Sennrich, Barry Haddow und Alexandra Birch (2016). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation,* Vol. 2: Shared Task Papers, 368-373, Berlin.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry und Maria Nadejde (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* 65–68,

Valencia.

Jörg Tiedemann (2017). The frustrating past, the exciting present and the bright future of (neural) machine translation, part 1. Youtube-Video. Vorlesung an der *Artificial Intelligence and Natural Language Conference,* St. Petersburg.

Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann und Joe Tebelskis (1991). Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal,* Vol. 2, 793–796, Toronto.

Lebenslauf

Persönliche Angaben

Mara Bertamini
Sonnmattweg 1
5620 Bremgarten
maratiziana.bertamini@uzh.ch

Schulbildung

seit 2015	Master in Multilinguale Textanalyse, Universität Zürich
2012-2015	Bachelor in Vergleichende Romanische Sprachwissenschaft, Computerlinguistik und Ethnologie, Universität Zürich

Berufliche und nebenberufliche Tätigkeiten

seit 2015	Swiss Life Holding AG, Zürich Assistentin Group HR Learning & Development
2012–2015	ABACUS Nachhilfeinstitut
2011-2017	Zurich Insurance Group, Zürich World Economic Forum, Davos Hotel Edelweiss, Davos Atelier West Architekten, Baden Link Institut, Zürich

A Modell-Konfigurationen


```
{
  "anneal_decay": 0.5,
  "batch_size": 80,
  "clip_c": 1,
  "datasets": [
    "/home/user/bertamini/data3/corpus.bpe.en",
    "/home/user/bertamini/data3/corpus.bpe.de"
  ],
  "dec_base_recurrence_transition_depth": 2,
  "dec_deep_context": false,
  "dec_depth": 1,
  "dec_high_recurrence_transition_depth": 1,
  "decoder": "gru_cond",
  "decoder_deep": "gru",
  "decoder_truncate_gradient": -1,
  "dictionaries": [
    "/home/user/bertamini/data3/corpus.bpe.en.json",
    "/home/user/bertamini/data3/corpus.bpe.de.json"
  ],
  "dim": 1024,
  "dim_per_factor": [
    512
  ],
  "dim_word": 512,
  "dispFreq": 1000,
  "domain_interpolation_inc": 0.1,
  "domain_interpolation_indomain_datasets": null,
  "domain_interpolation_max": 1.0,
  "domain_interpolation_min": 0.1,
  "dropout_embedding": 0.2,
  "dropout_hidden": 0.2,
  "enc_depth": 1,
  "enc_depth_bidirectional": 1,
  "enc_recurrence_transition_depth": 1,
  "encoder": "gru",
  "encoder_truncate_gradient": -1,
  "external_validation_script": null,
  "factors": 1,
  "finish_after": 10000000,
  "lrate": 0.0001,
  "map_decay_c": 0,
  "max_epochs": 5000,
  "maxibatch_size": 20,
  "maxlen": 50,
  "model_version": 0.1,
  "mrt_alpha": 0.005,
  "mrt_loss": "SENTENCEBLEU n=4",
  "n_words": 85000,
  "n_words_src": 85000,
  "objective": "CE",
  "optimizer": "adam",
  "patience": 10,
  "raml_reward": "hamming_distance",
  "raml_samples": 1,
  "raml_tau": 0.85,
  "reload_": false,
  "reload_training_progress": true,
  "sampleFreq": 10000,
  "saveFreq": 30000,
  "saveto": "/home/user/bertamini/model4/model.npz",
  "shuffle_each_epoch": true,
  "sort_by_length": true,
  "validFreq": 10000,
  "valid_batch_size": 80,
  "valid_datasets": [
    "/home/user/bertamini/data3/dev.bpe.en",
    "/home/user/bertamini/data3/dev.bpe.de"
  ],
}
```


Abbildung 8: Konfiguration BL-System

```
{
  "anneal_decay": 0.5,
  "batch_size": 80,
  "clip_c": 1,
  "datasets": [
     "/home/user/bertamini/data_sl/corpus.bpe.en",
    "/home/user/bertamini/data_sl/corpus.bpe.de"
  ],
  "dec_base_recurrence_transition_depth": 2,
  "dec_deep_context": false,
  "dec_depth": 1,
  "dec_high_recurrence_transition_depth": 1,
  "decoder": "gru_cond",
  "decoder_deep": "gru",
  "decoder_truncate_gradient": -1,
  "dictionaries": [
    "/home/user/bertamini/data3/corpus.bpe.en.json",
    "/home/user/bertamini/data3/corpus.bpe.de.json"
  ],
  "dim": 1024,
  "dim_per_factor": [
    512
  ],
  "dim_word": 512,
  "dispFreq": 1000,
  "domain_interpolation_inc": 0.1,
  "domain_interpolation_indomain_datasets": null,
  "domain_interpolation_max": 1.0,
  "domain_interpolation_min": 0.1,
  "dropout_embedding": 0.2,
  "dropout_hidden": 0.2,
  "enc_depth": 1,
  "enc_depth_bidirectional": 1,
  "enc_recurrence_transition_depth": 1,
  "encoder": "gru",
  "encoder_truncate_gradient": -1,
  "external_validation_script": null,
  "factors": 1,
  "finish_after": 10000000,
  "lrate": 0.0001,
  "map_decay_c": 0,
  "max_epochs": 1,
  "maxibatch_size": 20,
  "maxlen": 50,
  "model_version": 0.1,
  "mrt_alpha": 0.005,
  "mrt_loss": "SENTENCEBLEU n=4",
  "n_words": 85000,
  "n_words_src": 85000,
  "objective": "CE",
  "optimizer": "adam",
  "patience": 10,
  "raml_reward": "hamming_distance",
  "raml_samples": 1,
  "raml_tau": 0.85,
  "reload_": false,
  "reload_training_progress": false,
  "sampleFreq": 10000,
  "saveFreq": 30000,
  "saveto": "/home/user/bertamini/model4_da_ep1/model.npz",
  "shuffle_each_epoch": true,
  "sort_by_length": true,
  "validFreq": 10000,
  "valid_batch_size": 80,
  "valid_datasets": [
    "/home/user/bertamini/data3/dev.bpe.en",
    "/home/user/bertamini/data3/dev.bpe.de"
  ],
}
```


Abbildung 9: Konfiguration DA-System

B Ressourcen

Das domänenspezifische Trainingskorpus, die Testfiles, eigens kreierte Skripts zur Vorverarbeitung sowie die 100 als domänenspezifsche Termini definierten Adjektiv-Nomen-Kombinationen sind unter folgendem Link aufrufbar:
`https://drive.google.com/open?id=1jHHs4WkZfJpoOWFJ51T8g1c8GKLR-lkm`

Universität Zürich[UZH]

Institut für Computerlinguistik

Selbstständigkeitserklärung

Originalarbeit

Ich erkläre ausdrücklich, dass es sich bei der von mir im Frühjahrs-/Herbst-Semester 2018 ... an der Universität Zürich eingereichten schriftlichen Arbeit mit dem Titel

Domänenspezifische neuronale maschinelle Übersetzung

um eine von mir selbst und ohne unerlaubte Beihilfe sowie in eigenen Worten verfasste Originalarbeit handelt. Sofern es sich dabei um eine Arbeit von mehreren Verfasserinnen oder Verfassern handelt, bestätige ich, dass die entsprechenden Teile der Arbeit korrekt und klar gekennzeichnet und der jeweiligen Autorin oder dem jeweiligen Autor eindeutig zuzuordnen sind.

Ich bestätige überdies, dass die Arbeit als Ganzes oder in Teilen weder bereits einmal zur Abgeltung anderer Studienleistungen an der Universität Zürich oder an einer anderen Universität oder Ausbildungseinrichtung eingereicht worden ist noch inskünftig durch mein Zutun als Abgeltung einer weiteren Studienleistung eingereicht werden wird.

Verwendung von Quellen

Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit enthaltenen Bezüge auf fremde Quellen (einschliesslich Tabellen, Grafiken u. €.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos und nach bestem Wissen sowohl bei wärtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

Sanktionen

Ich nehme zur Kenntnis, dass eine Arbeit, welche zum Erwerb eines Leistungsnachweises verwendet wird und sich als Plagiat im Sinne des Dokuments *Erläuterung des Begriffs „Plagiat“* erweist, in leichten Fällen zu Notenabzug führt, in schweren Fällen mit Note 1 (eins) ohne Möglichkeit einer Überarbeitung bewertet werden kann und in ganz gravierenden Fällen die entsprechenden rechtlichen und disziplinarischen Konsequenzen nach sich ziehen kann (gemäss §§ 7ff der Disziplinarordnung der Universität Zürich sowie § 36 der Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich).

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben.

Name: Bertamini Vorname: Mara

Matrikelnummer: 11-919-347

Datum: 18.06.2018 Unterschrift: