Pre-reordering for Statistical Machine Translation of Non-fictional Subtitles

Magdalena Plamadă[1] **Gion Linder**[2] **Phillip Ströbel**[1] **Martin Volk**[1]

[1]Institute of Computational Linguistics
University of Zurich
Binzmühlestrasse 14
CH-8050 Zurich
{plamada, volk}@cl.uzh.ch
phillip.stroebel@uzh.ch

[2]SWISS TXT
Schweizerische Teletext AG
Alexander-Schöni-Strasse 40
CH-2501 Biel
gion.linder@swisstxt.ch

Abstract

This paper describes the challenges of building a Statistical Machine Translation (SMT) system for non-fictional subtitles. Since our experiments focus on a "difficult" translation direction (i.e. French-German), we investigate several methods to improve the translation performance. We also compare our in-house SMT systems (including domain adaptation and pre-reordering techniques) to other SMT services and show that pre-reordering alone significantly improves the baseline systems.

1 Introduction

The recent advances in Statistical Machine Translation (SMT) have drawn the interest of the language industry towards it. The main advantages of integrating automatic translations are both cost and time savings, since the translation efforts can be reduced to post-editing activities. Experiments for different topical domains (such as software localization, film subtitling or automobile marketing texts) reported time savings between 20% and 30% (Volk, 2008; Plitt and Masselot, 2010; Läubli et al., 2013). These success stories strengthen our motivation to build a SMT system specialized on non-fictional content (e.g. documentaries, informative broadcasts).

The challenge of this task lies in the desired translation direction, namely from French into German. As the target language is morphologically richer than the source language, we expect difficulties in generating grammatically correct output. This drawback can be overcome by means of hierarchical models (Huck et al., 2013), improved morphological processing (Cap et al., 2014) or models enriched with part-of-speech (POS) information (Stüker et al., 2011). Another known issue with translations into German is the word order (e.g. the long-range disposal of separable prefix verbs or composed tenses), which can result in missing verbs or verb particles in the translated output. A general solution when translating between languages with different word order is to reorder the source texts according to the word order in the target language, as suggested by Niehues and Kolls (2009).

In this paper we investigate how well these techniques can be applied for subtitles and we particularly focus on the problem of missing verbs. We show that handling this aspect alone improves the SMT performance. We furthermore discuss whether the SMT performance is good enough to be incorporated in the translation workflow of a subtitling company.

2 The proposed solution

2.1 Domain description

SWISS TXT provides multimedia solutions for the Swiss National Radio and Television association. The company includes a subtitling division, which is responsible for producing subtitles for the broadcasted TV shows in the Swiss national languages: German, French, Italian and Rumansh. The subtitles are localized for the region where the TV show is broadcast (e.g. in the German-speaking part of Switzerland subtitles are only displayed in German). In order to ensure the desired quality, this work is done manually.

In a small cooperation project, we investigated whether SMT can facilitate the translation process, with a special focus on translating the subtitles of a French TV news magazine (called TP[1]) into German. The magazine covers a variety of topics, such as politics, society, economy or history with both Swiss and international foci.

2.2 Reordering approach

Although the standard SMT training includes by default a reordering step, the model cannot handle long-distance verb components. Therefore we apply an additional reordering step on the French input during pre-processing (also called pre-reordering), in which we focus on verb "dependencies". Our approach is rule-based and makes the distinction between main and subordinate clauses, since the position of the German verbs differs from clause to clause. For example, in declarative main clauses the finite verb is in the second position, whereas in some interrogative and exclamatory sentences it is in initial position (verb first). And in some subordinate clauses it can take a clause-final position.

Our reordering rules are mostly based on POS tags, but sometimes they also include word lemmas. They are learned from a subset of the French treebank consisting of 12,500 sentences from the LeMonde newspaper (Abeillé and Barrier, 2004). We first tag and parse the French sentences[2] and identify the main and subordinate clauses. Subsequently we extract the POS sequences corresponding to main and subordinate clauses respectively, and calculate their frequency. The most frequent patterns are then manually analyzed and corresponding reordering rules are generated.

As an example, consider the French sentence *FR orig* (English: I hope that this will level off.) and the extracted reordering rule. In this case, the auxiliary verb *va* has to be placed in the end of the subordinate clause, in order to comply with the German word order (as in *FR reordered*).

FR orig J'*/CLS* espère*/V* que*/CS* ça*/PRO* va*/V* se*/CLR* stabiliser*/VINF ./PONCT*

PRO V CLR VINF → PRO CLR VINF V

FR reordered J'espère que ça se stabiliser va.

[1]Full name suppressed due to privacy concerns

[2]http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html

A frequency distribution of these patterns shows that there are a couple of reoccurring patterns and many tag sequences which are rare (in agreement with Zipf's law). The rule set in these experiments consists of 30 rules, which cover approximately 70% of the sentences in need for reordering.

3 SMT experiments

3.1 Data description

It is known that good SMT performance can be obtained with considerable amounts of similar training data. In our case, only 40 subtitle files of the TP magazine were available in both languages, since the TV show has only recently been broadcast in the German-speaking part. Therefore we had to make use of other parallel resources, as similar as possible to the texts we intend to translate. A brief description of the data sets follows:

In-domain data The dataset consists of the 40 comparable files[3] of the informative broadcast *TP*.

"Similar in-domain" data I [4] The dataset consists of TED talks transcriptions in German and French from the WIT3 corpus[5].

"Similar in-domain" data II [4] The dataset consists of subtitles of informative broadcasts with the same profile (called TV)[1].

Out-of-domain data The dataset consists of freely available subtitles from the OPUS OpenSubtitles corpus[6].

The size of the parallel data sets used for our SMT experiments is detailed in table 1. We report the number of sentences because we decided to train the system on whole sentences, since the bigger corpora (OPUS and TED) were already available in this format. For this purpose, TV and TP subtitles have also been merged into sentences. The development and the test data have been withheld from the in-domain corpus.

3.2 System description

The SMT systems are trained with the Moses toolkit, according to the guidelines on the official

[3]We call them comparable because not every German subtitle/sentence has a corresponding French one and vice versa.

[4]Non-fictional texts, written in a different style than the one to translate

[5]https://wit3.fbk.eu/

[6]http://opus.lingfil.uu.se/

Data set	Sentences	DE Words	FR Words
OPUS	3,326,000	20,635,000	20,853,000
TV	641,000	5,905,000	8,760,000
TED	137,000	2,166,000	2,881,000
TP	11,000	113,000	144,000
Dev set	1350	14,000	14,800
Test set	300	3,000	3,200

Table 1: The size of the German-French data sets

website, with the difference that we lowercase the data instead of truecasing it[7]. The model combinations (phrase table combination, language model interpolation) are generated with the tools available in the Moses distribution. The parameters of the global models are optimized through Minimum Error Rate Training (MERT) on an in-domain development set (Och, 2003). The translation performance is measured in terms of several evaluation metrics on a single reference translation using `multeval`[8].

Since the collected data sets are very heterogeneous, training a system on concatenated data did not make any sense because we would risk that bigger corpora overpower the small in-domain one. To avoid this, we make use of a common domain adaptation technique, namely mixture-modeling (Sennrich, 2012), and we apply it to both the translation and the language models. The components of the combined translation models have been trained independently on the corresponding parallel corpora (OPUS, TED etc.), whereas the language models are trained on the target side of these corpora.

The *Hierarchical* system is trained by the same principles, but uses hierarchical models instead of plain phrase-based models. Such models learn translation rules from parallel data by means of probabilistic synchronous context-free grammars and are able to handle languages with different word order. The *Improved* system uses mixed phrase-based models, but unlike the baseline system, the models are trained on reordered sentences. Reordering is performed during preprocessing and has been applied to training, development and test data alike. However, reordering only makes sense if the main clause and the subordinate ones are in the same translation unit. Since a common practice in subtitling is to separate subordinate clauses from the main clause (due to length restrictions), we had to join the subtitles in order for the reordering to be effective.

3.3 Results

The results of the SMT experiments are summarized in table 2. As expected, both the hierarchical and the improved systems outperform the baseline in the automatic evaluation, as reflected by all reported scores (BLEU, METEOR and TER).

System	BLEU ↑	METEOR ↑	TER ↓
Baseline	16.4	34.9	64.5
Hierarchical	17.1	35.3	64.2
Improved	17.4	35.9	63.9
Google Translate	14.3	30.3	68.7

Table 2: SMT results for French-German

However, the system trained on reordered sentences is slightly better than the hierarchical one, as the following example shows. We also compared our in-house systems against Google Translate (a large scale SMT system)[9] and we systematically score better. However, this effect can partially be attributed to the lexical choices, which are different from the reference, as the following example shows.

FR orig: -Rémy est loin d'imaginer ce qui va lui arriver .
Baseline: -Rémy ist nicht, was geschehen wird .
Hierarchical: Es ist nicht, was geschehen wird .
Improved: -Rémy ist weit weg, sich vorzustellen, was ihm geschieht .
Google: -Rémy hat keine Ahnung, was mit ihm geschehen wird .
DE ref: Rémy hat keine Vorstellung, was ihm bevorsteht .

The same happens with the *Improved* system, which generates an almost correct German sentence following the syntax from the original sentence (which is different from the reference). We also note that this output is better than what the rest of our in-house systems generate because the verbs are no longer missing and they are correctly placed according to the type of clause (main/subordinate). However, a better option would have been to translate the phrase *être loin d'imaginer* (EN: to be far from imagining) as a multiword unit, but our systems do not specifically handle these kind of phrases.

[7] `http://www.statmt.org/moses/?n=Moses.Baseline`
[8] `https://github.com/jhclark/multeval`
[9] `http://translate.google.com`

In order to assess the improvements from a translator's perspective, we conducted a small human evaluation experiment with one potential user. The purpose of the experiment was to judge the usefulness of the MT output in general, with respect to post-editing efforts. The test data consisted of a real subtitle file with no additional preprocessing (e.g. merging into sentences). According to his judgment, 33.5% of the subtitles can be used directly or with small corrections, 48.5% of the subtitles need improvements, but post-editing would still be faster than translating from scratch, whereas 18% of the subtitles require a retranslation. We consider these findings more insightful than the automatic scores, as they can be used to further improve our SMT system.

4 Conclusion

In this paper we have described our efforts of building a SMT system for translating French subtitles into German. This was particularly challenging since only a small in-domain corpus was available and thus different corpora (with different styles) had to be combined into a single system. We addressed this issue by applying mixture modeling, thus ensuring that Swiss-specific terms were preferred over alternative translations. For example, the French verb *évincer* (EN: to expel sb.) was consistently translated as *ausschaffen* (as learned from our in-domain corpus), instead of *ausschliessen* (as found in other corpora).

We have also shown how the translation quality can be improved by pre-reordering the input sentences. This preprocessing step used a set of POS-based rules extracted from a parsed French corpus. Although our approach focused on the correct placement of verbs depending on the clause type (main vs. subordinate), the system trained with reordered sentences gained 1 BLEU point on top of the baseline. This finding suggests that a more refined set of reordering rules will contribute to further improving translations. It is also conceivable to include morphological information (as suggested by other approaches) for the purpose of generating correct word forms.

We cannot help noticing that the obtained BLEU scores were still in a low range. We think that this was partially due to our test set, which often contained paraphrases instead of literal translations. On the other hand, the human evaluation showed a high acceptance rate of the MT output, since only 18% was assessed as unusable. This kind of output could be easily suppressed in a quality estimation post-processing step. This way we would only deliver translations in which our system is confident, allowing post-editors to save both time and efforts.

References

Abeillé, Anne and Nicolas Barrier. 2004. Enriching a French Treebank. *Proceedings of the Fourth Conference on Language Resources and Evaluation*.

Cap, Fabienne, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 579–587.

Huck, Matthias, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A Phrase Orientation Model for Hierarchical Machine Translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. 452–463.

Läubli, Samuel, Mark Fishel, Manuela Weibel, and Martin Volk. 2013. Statistical Machine Translation for Automobile Marketing Texts. *Proceedings of the Machine Translation Summit XIV*. 265–272.

Niehues, Jan and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. 206–214.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. 160–167.

Plitt, Mirko and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-editing in a Typical Localisation Context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Sennrich, Rico. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 539–549

Stüker, Sebastian, Kevin Kilgour, and Jan Niehues. 2011. Quaero Speech-to-Text and Text Translation Evaluation Systems. *High Performance Computing in Science and Engineering '10*. 529–542.

Volk, Martin. 2008. The Automatic Translation of Film Subtitles. A Machine Translation Success Story? *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*. 202–214.