Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses

Maja Popović
ADAPT Centre
Dublin City University
Ireland
maja.popovic@adaptcentre.ie

Abstract

We present a test set for evaluating an MT system's capability to translate ambiguous conjunctions depending on the sentence structure. We concentrate on the English conjunction "but" and its French equivalent "mais" which can be translated into two different German conjunctions. We evaluate all English-to-German and French-to-German submissions to the WMT 2019 shared translation task. The evaluation is done mainly automatically, with additional fast manual inspection of unclear cases.

All systems almost perfectly recognise the target conjunction "aber", whereas accuracies for the other target conjunction "sondern" range from 78% to 97%, and the errors are mostly caused by replacing it with the alternative conjunction "aber". The best performing system for both language pairs is a multilingual Transformer *TartuNLP* system trained on all WMT 2019 language pairs which use the Latin script, indicating that the multilingual approach is beneficial for conjunction disambiguation. As for other system features, such as using synthetic back-translated data, context-aware, hybrid, etc., no particular (dis)advantages can be observed.

Qualitative manual inspection of translation hypotheses shown that highly ranked systems generally produce translations with high adequacy and fluency, meaning that these systems are not only capable of capturing the right conjunction whereas the rest of the translation hypothesis is poor. On the other hand, the low ranked systems generally exhibit lower fluency and poor adequacy.

1 Introduction

Ambiguous words are often difficult to translate automatically, even by the current state-of-the-art neural machine (NMT) systems. Whereas NMT systems produce more fluent (grammatical and natural) translations than the previous state-of-the-art statistical phrase-based (PBMT) models, the semantic faithfulness of the translation to the original (adequacy) is still often problematic (Castilho et al., 2017; Klubička et al., 2018). Adequacy is even more problematic for ambiguous words which have two or more meanings depending on the context. Whereas the ambiguity of nouns, verbs and pronouns has been evaluated extensively in the recent years (Burchardt et al., 2017; Müller et al., 2018; Rios Gonzales et al., 2017, 2018), no results for conjunctions have been reported so far, and conjunctions can be ambiguous, too. It should be noted, though, that the conjunction ambiguity is more structural than lexical: it is mainly related to certain aspects of grammar involving the arrangement of words and word types. Therefore, the conjunction ambiguity is related more to fluency than to adequacy. The only work dealing with conjunctions and machine translation (Huang, 1983) explores conjunction scope for rule-based MT systems and does not address the ambiguity.

Our aim is to enable quantitative analysis of translating ambiguous conjunctions in a reproducible and semi-automatic way and to compare different types of systems in this respect. Our test sets for WMT 2019 are designed for the English ambiguous conjunction "but" and its French equivalent "mais", each of which can be translated into two different German conjunctions, "aber" or "sondern". The content is mainly based on general domain from subtitles (Tiedemann, 2012). Instead of comparing the translation hypotheses with a reference translation, we base the evaluation on the presence or absence of the correct conjunction in the target language. For unclear cases (about 1% of segments), manual inspection is carried out. We report results on all English→German and French→German submissions to the WMT 2019 shared translation task.

In addition to German, the test sets can be used for any target language which has these two variants of the conjunction ”but” (for example Spanish or Croatian).

2 German equivalents of ”but”/”mais”

The English coordinating conjunction ”but” and its French equivalent ”mais” are ambiguous when translated into certain target languages such as German. In German, there are two possible variants, ”aber” and ”sondern”. ”Aber” can be used after either a positive or a negative clause. On the other hand, ”sondern” is only used after a negative clause when expressing a contradiction. The first clause in the sentence must contain a negation marker, and the second part of the sentence must contradict the first part of the sentence.

Three examples can be seen in Table 1. The sentences on the left have the same context, same or similar meanining, and contain similar words as the sentences on the right. Nevertheless, the conjunction ”but” in all sentences on the left should be translated as ”aber” and in those on the right as ”sondern”. This illustrates the statement from the previous section about the structural nature of conjunction ambiguity.

Generally, sentences with ”aber” can be found more frequently in the data. Table 2 presents the distribution of the two types of sentences in the WMT 2019 News Commentary training corpus. In addition, it can be noted that both types of sentences occure rarely in the News corpus (less than 4% in total).

3 Test sets

3.1 Preparation

The test sets are generated semi-automatically using the bilingual subtitles corpora[1] according to the following requirements: (i) include only short segments (up to 20 words) (ii) remove all noise (iii) avoid complex words and rare name entities which could introduce additional effects.

First step was to extract all short segments containing the desired conjunctions in the source (English and French) and the target (German) language, and the second step was manual elimination or rephrasing complex and noisy parts. In this way, about 1000 sentences for each of the source languages were prepared, containing about 800 instances of ”sondern” and 200 instances of ”aber”. Since our preliminary experiments shown that the sentences requiring ”aber” are less difficult for MT systems, we concentrate more on the performance for the conjunction ”sondern”.

A detailed corpus statistics is presented in Table 3. It can be seen that the segments are relatively short, and the vocabulary size relatively low – the vocabulary size of the standard English test set from WMT 2018 is more than double, about 5000 distinct words, and the average sentence length is 22.5. Apart from this, it can be seen that the average segment length of the easier ”aber” instances is slightly lower.

It should be noted that, although the basis for the generation of the test sets was a bilingual corpus, the resulting test sets do not contain any reference translations. The reason for this is twofold: on the one hand, bilingual manual filtering of noisy and complex content would be very time and resource consuming. On the other hand, reference translations are not really needed – since we are interesting only in conjunction disambiguation, checking the conjunction in the translation hypothesis is sufficient and it can be carried out without a reference translation.

3.2 Evaluation

The vast majority of checks is performed automatically, however for a small number of sentences (usually 1-2%) a manual inspection is needed. For each sentence, there are four possible outcomes of the automatic evaluation:

- only the correct conjunction is found
  $\Rightarrow$ correct
- only the opposite conjunction is found
  $\Rightarrow$ incorrect
- both conjunctions are found
  $\Rightarrow$ manual inspection
- none of the two conjunctions is found
  $\Rightarrow$ manual inspection

Manual inspection is carried out in the following way: if the structure of a sentence with additional or without any conjunctions is correct, then the sentence is considered correct. All errors which are not related to the conjunction are ignored, both by automatic and by manual evaluation.

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php

”aber”	”sondern”
You’re apologizing to me, but you should apologize to her.	Don’t apologize to me, but to her.
The child wanted to go to the park, but we went home.	The child didn’t want to go home, but to the park.
You should never speak but you can write.	You should never speak but only write.

Table 1: Examples of difference between the two German conjunctions.

lang. pair	aber	sondern
En-De	8230 (2.4%)	4389 (1.3%)
Fr-De	5498 (2.1%)	3369 (1.3%)

Table 2: Distribution of sentences requiring each of the two German conjunctions in the News Commentary training corpus for WMT 2019: number of sentences and percentage in the whole corpus.

4 MT Systems

4.1 English-to-German

All English-to-German systems are trained on the constraint data except *en-de-task* and *PROMT-NMT*. For the *en-de-task* system, as well as the *Microsoft-doc/sent level* systems, no additional information is available.

All other systems are based on the Transformer architecture, and *UCAM* uses the phrase-based approach too, thus being the only hybrid system.

All systems used BPE[2] segmentation except *eTranslation* which used SentencePiece[3] segmentation.

MSRA.MADL, *TartuNLP* and *UdS-DFKI* were trained only on natural parallel data, whereas all other systems used synthetic back-translated data, too. *JHU*, *NEU* and *UCAM* performed back-translation more than once.

The *LMU* and *UdS-DFKI* systems are context aware, *UdS-DFKI* being coreference aware.

MSRA.MADL used multi-agent dual learning (MADL)[4].

The only multilingual system is *TartuNLP*, one and the same Transformer system trained on all WMT language pairs which use Latin script.

4.2 French-to-German

All French-to-German systems are based on the Transformer architecture and used the constrained data.

All systems used BPE units except *eTranslation* which used SentencePiece units.

MSRA.MADL and *TartuNLP* are trained only on natural parallel data, whereas *eTranslation, LIUM* and *MLLP-UPV* used additional synthetic back-translated data.

MSRA.MADL again used multi-agent dual learning (MADL).

TartuNLP is again the only multilingual system, the same one used for the English-to-German task.

5 Results

The results are presented in Table 4 in the form of percentage of sentences automatically identified as correct (”aut.”), identified as correct after both automatic check and manual inspection (”full”), and automatically identified as incorrect because the source conjunction is translated into the opposite conjunction (”opposite”). The systems are ranked by the full accuracy of the conjunction ”sondern”.

5.1 General observations

Generally, the same tendencies are observed for both language pairs.

First of all, in can be noted that the results of our preliminary experiments mentioned in Section 2 are confirmed on the large scale: translating sentences requiring the conjunction ”aber” is not problematic for any of the systems: the percentage of correct sentences is 100%, or in the worst cases, close to 100%, for both language pairs and all systems.

As for the ”difficult” conjunction ”sondern”, the majority of the systems translates it correctly in 90-95% of cases, and the predominant problem for the rest is translating it as ”aber” (5-10%). Other types of errors are found in only very small number of cases (for example, parts of the sentences left untranslated, or completely incorrect sentence structure).

For the sentences with both conjunctions or without any of the two conjunctions, manual in-

[2]https://github.com/rsennrich/subword-nmt
[3]https://github.com/google/sentencepiece
[4]https://openreview.net/pdf?id=HyGhN2A5tm

source language	target conjunction	number of sentences	number of running words	vocabulary size	average sent. length
English	all	1066	13655	2252	12.8
	”sondern”	858	11058	2043	12.9
	”aber”	208	2597	560	12.5
French	all	1010	12963	2162	12.8
	”sondern”	806	10478	1823	13.0
	”aber”	98.1	2485	673	12.2

Table 3: Statistics of the test sets: number of sentences, number of running words, vocabulary size and average sentence length.

language pair	system	”sondern” correct aut.	”sondern” correct full	”sondern” opposite (”aber”)	”aber” correct aut.	”aber” correct full	”aber” opposite (”sondern”)
En→De	TartuNLP	97.2	97.3	2.7	98.6	99.0	1.0
	NEU	96.1	96.1	3.8	100	100	0
	HelsinkiNLP	95.3	95.6	4.3	99.0	99.5	0
	MSRA.MADL	94.5	94.6	5.1	99.5	99.5	0
	dfki-nmt	94.0	94.6	5.2	99.0	99.5	0.5
	online-A	94.3	94.4	5.3	99.0	99.0	1.0
	eTranslation	94.0	94.3	5.5	100	100	0
	Microsoft-sent-level	93.8	93.9	6.1	99.5	100	0
	Facebook-Fair	93.6	93.7	6.2	100	100	0
	Microsoft-doc-level	93.6	93.6	6.3	100	100	0
	UdS-DFKI	92.8	92.8	6.7	99.0	99.0	0
	LMU	91.6	91.8	7.8	95.2	95.7	1.0
	UCAM	91.7	91.7	8.2	99.0	99.0	1.0
	JHU	91.4	91.7	8.2	100	100	0
	MLLP-UPV	91.0	91.2	8.4	100	100	0
	online-Y	90.3	90.3	9.6	99.5	99.5	0.5
	PROMT-NMT	89.4	89.4	9.9	100	100	0
	online-B	88.8	89.4	10.2	99.0	99.5	0
	online-G	89.0	89.2	10.7	100	100	0
	online-X	86.0	86.0	13.7	99.5	99.5	0.5
	en-de-task	78.2	78.2	21.3	95.2	95.7	3.4
Fr→De	TartuNLP	96.9	96.9	3.1	97.5	98.5	0.5
	eTranslation	93.0	93.4	6.6	100	100	0
	online-G	87.6	93.4	6.7	100	100	0
	MSRA.MADL	93.2	93.3	6.7	100	100	0
	online-A	88.5	92.8	6.7	100	100	0
	MLLP-UPV	92.0	92.4	7.4	99.5	99.5	0.5
	LIUM	91.3	91.7	8.3	100	100	0
	online-B	87.3	89.7	10.5	100	100	0
	online-Y	67.9	88.7	10.5	100	100	0
	online-X	86.8	86.8	13.2	100	100	0

Table 4: Percentage of correct conjunctions retrieved automatically and by full evaluation, and percentage of opposite conjunctions.

source:	*However*, this is not Agnes, **but** her daughter.
output:	Das ist *aber* nicht Agnes, **sondern** ihre Tochter.
source:	The time, *however*, is not thirty years ago, **but** now.
output:	Die Zeit is *aber* nicht dreissig Jahre her, **sondern** jetzt.

Table 5: Examples of correct translations with both German conjunctions.

spection is carried out. For English-to-German systems, only a small number of sentences fall into these two categories, so that manual inspection has no or very little effect on ranking. For four "online" French-to-German systems, online-A, -B, -G and -Y, however, a larger number of sentence without conjunctions is found.

Both conjunctions: Manual inspection revealed that this is not problematic: it can happen if "however", "yet" or similar word which can be translated as "aber" is present in the source sentence. Two examples can be seen in Table 5.

No conjunctions: For the English source, it can happen for a small number of sentences with structure "not only X, but Y, too", whereas for the French source a number of other sentence structures was paraphrased, too. Some of these paraphrased translations are perfect, whereas some of them are not as fluent as they would be if the construction with conjunction were used, but are nevertheless considered as correct. Two examples can be seen in Table 6.

source	Ce n'est pas un robot, **mais** un humain.
source (en gloss)	It is not a robot, **but** a human.
output	Er ist kein Roboter, er ist ein Mensch.
output (en gloss)	He is not a robot, he is a human.
source	Ce n'taient pas des mots, **mais** des actes.
source (gloss en)	It were not the words, **but** the deeds.
output	Es waren keine Worte, es waren Taten.
output (en gloss)	It was not words, it was deeds.

Table 6: Examples of correct translations without any of the two German conjunctions (mostly occuring in French-to-German systems).

5.2 Differences between the systems

The first and very interesting observation is that the best performing system for both language pairs is the multilingual *TartuNLP* system. The advantage of a multilingual system is probably its ability to get a signal for different structures from many languages, so that the information about different variants of the target conjunction necessary for different source sentence structures is better captured.

As for other system features, no particular differences can be spotted. For example, the best system *TartuNLP* is trained only on natural parallel data, the other system without back-translation *MSRA.MADL* performed very well, one system using multiple back-translation *NEU* is ranked as second and two other such systems *JHU* and *UCAM* in the middle, so no (dis)advantage of synthetic parallel data can be observed. Furthermore, two context-aware English-to-German systems *LMU* and *UdS-DFKI* as well as the hybrid *UCAM* system are ranged in the middle, thus no clear (dis)advangates of either of the approaches can be noted.

Qualitative analysis of overall performance

In order to check whether the best ranked systems maybe produce generally poor translations and only capture the conjunctions correctly, as well as other way round (maybe the lowest ranked systems produce fluent and adequate translations), we carried out a manual qualitative inspection of five highest and five lowest ranked hypotheses. The most important finding is that the best ranked systems produce decent translations both in terms of adequacy and fluency, meaning that these systems are not only capable of choosing the right conjunction while generating poor translations. As for the low ranked systems, they all have much lower fluency and adequacy, especially the lowest ranked *en-de-task* system with very low adequacy and a number of non-existing words.

Of course, to draw stabler conclusions, a systematic quantitative analysis of correlation between conjunction disambiguation and adequacy/fluency should be carried out in future work.

6 Conclusions

We present a targeted evaluation of 21 English-to-German and 10 French-to-German MT systems regarding their performance in lexical choice for ambiguous source conjunction ”but”/”mais”. We observe that all systems almost perfectly recognise the target conjunction ”aber”, whereas accuracies for the other target conjunction ”sondern” range from 78% to 97%, and the errors are mostly caused by replacing it with the alternative conjunction ”aber”.

The best performing system on the ”difficult” target variant ”sondern” for both source languages is based on the multilingual transformer model trained on all WMT language pairs using Latin script. The advantage of a multilingual system might be a better ability to learn the relation between different sentence structures and corresponding conjunctions. Apart of this, there are no other clear differences between the systems.

Qualitative analysis of translation hypotheses shown that highly ranked systems generally produce translations with high adequacy and fluency, meaning that they are not only capable of capturing the right conjunction whereas the rest of the translation hypothesis is poor. On the other hand, the low ranked systems generally exhibit lower fluency and poor adequacy. Quantitative analysis of correlation between the conjunction disambiguation and overall performance should be a part of future work.

The current study is focused on only one ambiguous conjunction and only one target language. In future, we plan to extend the test set with more conjunctions (and variants), and possibly, to more language pairs.

Acknowledgments

This research was supported by the ADAPT Centre for Digital Content Technology at Dublin City University, funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

References

Aljoscha Burchardt, Vivien Macketanz, Jonathan Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of pbsmt and nmt using professional translators. In *Proceedings of MT Summit XVI*, pages 116–131.

Xiuming Huang. 1983. Dealing with Conjunctions in a Machine Translation Environment. In *Proceedings of the 1st Conference on European Chapter of the Association for Computational Linguistics (EACL 1983)*, pages 81–85, Pisa, Italy.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: A case study on english to croatian. *Machine Translation*, 32(3):195–215.

Mathias Müller, Annette Rios Gonzales, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, pages 61–72, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the 2nd Conference on Machine Translation (WMT 2017)*, pages 11–19, Copenhagen, Denmark.

Annette Rios Gonzales, Mathias Mller, and Rico Sennrich. 2018. The word sense disambiguation test suite at wmt18. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey.