Investigating gender adaptation for speech translation

Rachel Bawden Guillaume Wisniewski Hélène Maynard
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
firstname.lastname@limsi.fr

ABSTRACT

In this paper we investigate the impact of the integration of context into dialogue translation. We present a new contextual parallel corpus of television subtitles and show how taking into account speaker gender can significantly improve machine translation quality in terms of BLEU and METEOR scores. We perform a manual analysis, which suggests that these improvements are not necessary related to the morphological consequences of speaker gender, but to more general linguistic divergences.

RÉSUMÉ

Étude de l'adaptation au genre du locuteur pour la traduction de la parole

Dans cet article nous évaluons l'impact de la prise en compte du contexte dans la traduction de dialogues. Nous introduisons pour cela un nouveau corpus parallèle, issu des sous-titres de séries télévisées, comportant de nombreuses informations contextuelles. Nous montrons comment la prise en compte du genre du locuteur permet d'améliorer significativement la qualité de la traduction automatique en termes de score BLEU et METEOR. Une analyse manuelle montre toutefois que ces gains ne sont pas nécessairement liés aux conséquences morphologiques du genre du locuteur, mais à des différences linguistiques plus générales.

MOTS-CLÉS : Traduction de la parole, TA, genre, adaptation, corpus parallèle.

KEYWORDS: Speech translation, SMT, gender, adaptation, parallel corpus.

1 Introduction

Journalistic and planned texts have long been the focus of attention in the Statistical Machine Translation (SMT) community largely due to the availability of large parallel corpora from parliamentary debates and the relative lack of data in other domains. Over the last decade, more and more emphasis has been placed on speech translation (Hardmeier, 2012), made possible thanks to the availability of speech-like parallel corpora such as the Ted talks and parallel subtitles (Lison & Tiedemann, 2016).[1] The translation of speech-like texts is a challenge for standard SMT systems (trained on journalistic or parliamentary corpora) due to the very different nature of the texts; linguistically, spoken sentences are often ungrammatical, incomplete and contain a greater degree of lexical diversity, but they must also be situated within a dialogue context, in which prosody, common ground and speaker information can offer the only ways of providing a correct translation. One such example is that of French gender agreement, in utterances such as "I am tired", translated as "Je suis fatigué**e**" for a female speaker and

[1] Although not strictly spontaneous speech, subtitles offer an approximation of speech and dialogue, and despite often containing a number of non-literal translations, have the advantage of existing in large quantities for multiple language pairs.

"Je suis fatigué" for a male speaker, for which speaker gender is the only possible way of determining the correct adjectival suffix.[2]

In this work, we aim to evaluate the impact of integrating external contextual information on translation quality. We present two contributions: (i) the creation of a contextualised parallel corpus of spontaneous dialogues, taken from television subtitles adapted from the TVD dataset (Roy *et al.*, 2014) and (ii) exploratory experiments on adapting translation systems to one example of contextual information: the gender of the speaker. We provide preliminary results for the translation of English to French subtitles, using automatic evaluation metrics, as well as a manual evaluation of improvements.

Statistical approaches to machine translation heavily depend on training data used, and therefore adapting systems to specific corpora is a very natural move. The choice of data used to tune model parameters is also highly important, as shown by Pecina *et al.* (2012) when tuning an out-of-domain model using in-domain data. Gender-dependent modelling is a common technique used in speech recognition to adapt models to acoustic differences between men and women's speech (Wahlster, 2000). However model adaptation is not restricted to speech recognition; Kübler *et al.* (2010) use dialogue act tags to improve a PoS tagger, and in machine translation, adaptive modelling has been used to create topic-dependent (Foster & Kuhn, 2007) and sentence-type-dependent (Finch *et al.*, 2009) models by training and tuning on class-partitioned data.

Our paper is organised as follows. We first describe our TV series corpus production process (Section 2). In Section 3 we present the baseline models, based on pre-existing corpora. In Section 4 we suggest and evaluate a number of adaptations of the baseline models to gender-specific data. Finally, Section 5 provides a more detailed analysis of these changes through a contrastive manual evaluation between gender-specific and baseline models.

2 *The Big Bang Theory* reproducible corpus

The TVD corpus: We focus on the translation from English to French of a small but contextually rich subtitle corpus: the first two seasons of the American television series *The Big Bang Theory*. The dataset is generated using the TVD plugin (Roy *et al.*, 2014), developed to provide reproducible datasets, exploiting visual, auditory and textual data, directly extracted from DVDs and freely available web sources.[3] This is an important addition to primarily textual resources such as OpenSubtitles, which provide little extra contextual information. Amongst the numerous contextual elements available (speech turns, audio signal and images), we choose, in this preliminary work, to focus on and integrate one aspect of this contextual information into translation - the gender of the speaker.

Extracting an English-French parallel subtitle corpus: To use the corpus for the purpose of translation, we first extracted an enriched parallel corpus from *The Big Bang Theory* TVD plugin. The text is taken from the official OCR-extracted subtitles,[4] since, unlike the manual transcriptions, they exist in multiple languages and are official translations. Following some correction of OCR errors using a small error lexicon and some manual rules, we align the French and English subtitles using their timestamps. Since there is not always a one-to-one mapping between French and English subtitles, and a perfect temporal alignment is not always present, heuristics are used to concatenate

[2]This information is not available in the source sentence and so is only attainable through knowledge of speaker gender.

[3]Due to copyright restrictions, the corpus cannot be distributed. However it can be easily reproduced using the plugin, once the DVD has been purchased.

[4]Using the Tesseract software (Smith, 2007) and VobSub2SRT (`https://github.com/ruediger/VobSub2SRT`)

subtitles where necessary in order to create parallel subtitle blocks.

Assigning gender to subtitles: Gender is assigned to subtitles by manually mapping speaker names to their corresponding gender. Speaker identities are provided by manual transcripts,[5] which are automatically aligned to the audio signal (Bredin *et al.*, 2014). We then transfer speaker identities to the subtitles, based on transcript and subtitle timestamps. We leave the task of using automatically predicted gender to future work; here we only consider reference genders to test our hypothesis.[6]

Division into train, development and test sets: We divide this corpus into three datasets: BBT-train (the first 40 episodes), used to train translation and language models, BBT-dev (the next 6 episodes), used for tuning and BBT-test (the last 6 episodes), used for evaluation. We partition each set into two subsets, one for each gender. Basic corpora statistics can be found in Table 1. Note that there is a strong class imbalance towards male speakers, who produce approximately ¾ of all test sentences. Subtitles corresponding to female speakers are also on average shorter than those for male speakers, and the percentage of out-of-vocabulary tokens compared to the two subtitle corpora (OPENSUBTITLES and BBT-train) is much smaller for female than male speakers, perhaps indicating a less heterogeneous use of vocabulary. Note that these characteristics are corpus-specific.

3 Baseline systems: testing pre-existing corpora

The first step in translating new data is to see how existing data fares for training a standard translation system. We provide baseline models using two pre-existing parallel corpora: EUROPARL (Koehn, 2005) and, more adapted to our domain, (though less commonly used) the film and television subtitle corpus OPENSUBTITLES (Lison & Tiedemann, 2016).[7] As this corpus is very large, we used the Modified Moore-Lewis (MML) Filtering algorithm (Axelrod *et al.*, 2011) to keep only the 8% of OPENSUBTITLE sentences most similar to BBT-train. We hereafter refer to this set as OpenSubs-mml.

A comparison of corpora is given in Table 1. One side-effect of filtering is the drop in average length between the corpus before and after filtering, most likely an effect of the fact that longer sentences are more different from each other in the two subtitle corpora. Despite this difference, OpenSubs-mml appears to be the most adapted corpus vocabulary-wise to our BBT data, resulting in the fewest out-of-vocabulary tokens. Note also the relative sentence lengths of source and target sentences. Whereas it is generally the case that French translations are longer than the corresponding English ones (as is the case with EUROPARL), the opposite is seen for the subtitle corpora, most probably linked to a shortening during subtitling due to the use of less literal translations and space constraints.

Translation setup: We use the Moses toolkit (Koehn *et al.*, 2007) for translation. Alignments are computed over all three training sets. All language models are 4-gram models with Kneser-Ney smoothing. Parameters are estimated on BBT-dev using `kbmira` to optimise the BLEU score. For each corpus, a separate phrase table and language model are produced. Multiple language and translation models are combined using the default Moses model combination approach, and are assigned weights during the tuning step.[8]

[5] `http://bigbangtrans.wordpress.com`

[6] Gender identification is a standard part of speaker diarisation systems (e.g. Barras *et al.*, 2006).

[7] We remove all episodes from *The Big Bang Theory* from this second corpus to avoid any bias.

[8] For example, the model EUROPARL+BBT-train contains two phrase tables (with the *either* strategy implemented in Moses), one for each corpus, and two language models, one for each corpus. A single reordering model is used for each system, based on the largest corpora used for the system.

Corpus	# sents.	Ave. sent. len. (en)	Ave. sent. len. (fr)	OpenSubs-mml % OOVs	EUROPARL % OOVs	BBT-train % OOVs
BBT-train	9,592	9.0	8.4	2.1	4.4	0
BBT-train$_m$	7,462	9.0	8.5	15.4	23.0	0
BBT-train$_f$	1,941	8.9	8.1	8.5	17.4	0
BBT-dev	2,089	9.1	8.1	2.1	4.0	7.3
BBT-test	1,941	9.2	8.2	1.9	4.0	7.2
BBT-test$_m$	1,438	9.4	8.4	2.3	4.1	8.0
BBT-test$_f$	354	8.8	8.0	1.1	4.0	4.7
EUROPARL	1,969,197	27.1	30.0			
OPENSUBTITLES	27,737,442	9.4	8.9			
OpenSubs-mml	2,218,997	5.9	5.4			

Table 1: Corpus statistics

	BBT-dev		BBT-test	
Model data	BLEU	METEOR	BLEU	METEOR
BBT-train	13.64	0.327	13.95	0.326
EUROPARL	14.11	0.328	13.99	0.326
OpenSubs-mml	23.20	0.419	23.69	0.432
EUROPARL + BBT-train	16.86	0.362	16.82	0.368
OpenSubs-mml+ BBT-train	23.39	0.419	24.09	0.434
OpenSubs-mml+EUROPARL	24.55	**0.430**	**24.74**	**0.441**
OpenSubs-mml+EUROPARL + BBT-train	**24.64**	0.428	24.56	0.437

Table 2: Automatic evaluation of baseline models.

Results: We evaluate the different models using two metrics: BLEU, and METEOR.[9] Results are shown in Table 2. As the models are tuned with BLEU, we judge the best model combination to be the one with the highest scores with the second metric METEOR. This model, OpenSubs-mml+EUROPARL also produce the highest scores for both metrics on the test set. It is unsurprising that both the BBT-train and EUROPARL models generalise poorly; the first lacks coverage because of its small size and the second because it is ill-adapted to speech-like data. However adding EUROPARL to OpenSubs-mml, by far the best-adapted dataset, does improve the scores of the latter. Somewhat surprising is the fact that adding BBT-train does not further improve results, and even degrades them slightly, most likely due to overfitting, as indicated by the BLEU scores on the development set.

4 Gender-based adaptative modelling

Taking our best baseline system (OpenSubs-mml+EUROPARL), we propose a series of adaptations to take speaker gender into account : (i) changing the tuning data, (ii) adding a gender-specific phrase table, (iii) adding a gender-specific language model, and (iv) adding both a gender-specific phrase table and language model. The additional gender-specific models are estimated using BBT-train sentences uttered by either male or female speakers. We distinguish three types of tuning data: 'all' corresponding to the entire BBT-dev, 'female' to sentences by female speakers and 'male' by male speakers. We test each of the models individually on the female and male BBT-test data.

[9] METEOR scores range from 0 to 1; the higher the score, the better the translation.

The results (Table 3) show that exploiting speaker gender is useful for translation, with improvements possible for both male and female speakers as shown on the test set. All adaptations provide some improvement in at least one configuration when compared to the baseline score (first row). The highest score seen for the male test set was the combination of a specific language model, a specific translation model and BBT-dev$_m$ for tuning (+LM$_m$+TM$_m$/male), with an improvement of +0.17 BLEU. The improvement between the baseline and contextualised model was greater for female speakers, with an improvement of +1.09 BLEU for the configuration (+LM$_f$/female). When the gender-adapted models are used, (and the baseline for unknown genders), we see an improvement in scores for BBT-dev (24.61 BLEU, 0.433 METEOR) and BBT-test (25.11 BLEU, 0.444 METEOR).

		BBT-test$_m$		BBT-test$_f$	
Model adaptation	Tuning data	BLEU	METEOR	BLEU	METEOR
(i) Choice of the tuning set					
∅	all	23.91	0.434	25.16	0.450
∅	male	**24.09**	**0.438**	**25.72**	0.450
∅	female	23.67	0.431	25.22	0.446
(ii) Addition of a gender-specific language model					
+LM$_m$	all	**24.17**	**0.436**	24.80	0.447
+LM$_f$	all	23.35	0.430	24.13	0.443
+LM$_m$	male	23.92	0.435	25.39	0.448
+LM$_f$	female	23.97	0.444	**26.25**	**0.459**
(iii) Addition of a gender-specific translation model					
+TM$_m$	all	23.94	0.436	25.12	0.443
+TM$_f$	all	23.71	0.432	24.93	0.447
+TM$_m$	male	23.84	0.433	25.25	0.443
+TM$_f$	female	23.54	0.432	25.38	0.450
(iv) Addition of a gender-specific language model and translation model					
+LM$_m$+TM$_m$	all	24.06	0.434	25.36	0.449
+LM$_f$+TM$_f$	all	23.60	0.431	24.55	0.444
+LM$_m$+TM$_m$	male	**24.18**	**0.436**	**25.69**	**0.451**
+LM$_f$+TM$_f$	female	22.64	0.422	24.91	0.441

Table 3: Translation performance after adaptation of the OpenSubs-mml+EUROPARL model

5 Discussion and analysis of improvements

We base our discussion on the differences between baseline predictions (∅/all) and those from the model that gave the greatest improvements for each gender: (LM$_m$+TM$_m$/male) and (LM$_f$/female).

Given the language pair (English to French), one type of error that we could hope to have corrected is that of gender agreement, particularly for adjectives and past participles, in sentence such as "I am happy", translated as "je suis contente" for a female speaker but "je suis content" for a male speaker. However given the small size of our dataset, the cases of this phenomenon are few; we manually identify 11 cases in the female test set which, given the lexical choice, could have resulted in a correction of gender, only one of which actually resulted in a correction.[10]

[10]We even identify a case of reported speech uttered by a female speaker, in which the gender was erroneously corrected:

So what are the improvements down to? As for many statistical systems, the improvements appear to be diverse and specific to the data used. To better understand the differences found between the baseline predictions and those from the gender-adapted models, we manually compared the quality of these two sets of translations and annotated their differences (See Table 4).[11]

The most common differences for both genders were in lexical choices, followed by additions and deletions. A change in lexical choice was more often associated with an improved translation than a degraded one (38% vs. 31% for both genders). However the change most linked to an improvement was addition, and conversely, the change most linked to a degradation was deletion; for male speakers, 73% of sentences whose only difference was an addition were improved (82% for females), and 93% of sentences whose only difference was a deletion were degraded (60% for female).

These observations suggest that the difference in BLEU score might result from differences in sentence length; the BLEU metric heavily penalises translation hypotheses that are shorter than the reference and, as shown in Table 1, female utterances are on average shorter than male utterances. It turns out that the baseline model produces translations that are 99.3% shorter than the reference translations for male speakers and 97% shorter for female speakers, whereas the adapted models produce translations 99.8% shorter for male speakers and 98.8% shorter for female speakers. The heightened improvements linked to addition for female speakers and the decreased effect of deletion compared to male speakers may be explained by the class imbalance in the data. The generic data used to tune the baseline model contains three times as much male data as female data, and therefore the baseline model is biased towards male utterances. Any improvements are therefore lessened with respect to the improvements that can be achieved by adaptation for the more sparse female tuning data.

	Number of differing translations				% of differing sentences that contain a change						
	Total	Better	Worse	Neutral	Add.	Del.	Reord.	Lex. choice	Tense	Gdr. agr.	Tu/Vous
Male	200	76	64	60	28.5	17.5	14	58	4	0.5	5
Female	114	50	32	32	35	15	9.5	55	5	3.5	9.5

Table 4: Manual analysis of differences between baseline and the best gender-adapted models.

6 Conclusion

We have shown that it is possible to improve manual and automatic evaluation scores when testing our subtitle corpus on baseline and gender-adapted models. A manual evaluation indicates that these preliminary results do not yet enable us to link these improvements to gender-specific linguistic phenomena such as gender agreement. Improvements appear to be due to other specificities of the datasets, such as average sentence length, and the higher gains for female speakers are almost certainly linked to class imbalance in the data. Further investigations will be needed to fully understand our results. We also intend to extend the work to a larger and more balanced dataset, in order to see whether further improvements can be made. Automatically identifying speaker gender will also enable us to forgo the need for manual transcriptions. We will also turn to other types of contextual information, such as the audio signal and speech turns, to improve translation using dialogue structure.

"The man said 'I am a physicist"', translated as "L'homme a dit 'je suis physicienne"', with a feminine suffix *-ienne*.

[11] We annotated all 114 of the female utterances that differed between the baseline and adapted model, and we randomly selected 200 of the 523 differing sentences for male speakers (of a total of 1,438 test sentences).

References

AXELROD A., HE X. & GAO J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, p. 355–362, Edinburgh, Scotland, UK.

BARRAS C., ZHU X., MEIGNIER S. & GAUVAIN J. L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, **14**(5), 1505–1512.

BREDIN H., ROY A., PÊCHEUX N. & ALLAUZEN A. (2014). "Sheldon speaking, bonjour!" - Leveraging Multilingual Tracks for (Weakly) Supervised Speaker Identification. In *Proceedings of the 22nd ACM International Conference on Multimedia (ACM-MM '14)*, p. 137–146, Orlando, USA.

FINCH A., SUMITA E. & NAKAMURA S. (2009). Class-Dependent Modeling for Dialog Translation. *IEICE Transactions on Information and Systems*, **92**(12), 2469–2477.

FOSTER G. & KUHN R. (2007). Mixture-Model Adaptation for SMT. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT '07)*, p. 128–135, Prague, Czech Republic.

HARDMEIER C. (2012). *Discourse in statistical machine translation. a survey and a case study*. PhD thesis, Uppsala University, Uppsala, Sweden.

KOEHN P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, p. 79–86, Phuket, Thailand.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Interactive Poster and Demonstration Sessions (ACL '07)*, p. 177–180, Prague, Czech Republic.

KÜBLER S., SCHEUTZ M., BAUCOM E. & ISRAEL R. (2010). Adding context information to POS tagging for dialogues. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT '10)*, p. 115–126, Tartu, Estonia.

LISON P. & TIEDEMANN J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC '16)*, p. 923–929, Portorož, Slovenia.

PECINA P., TORAL A. & VAN GENABITH J. (2012). Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. In *Proceedings of 24th Conference on Computational Linguistics (COLING '12)*, p. 2209–2224, Mumbai, India.

ROY A., GUINAUDEAU C., BREDIN H. & BARRAS C. (2014). TVD: A Reproducible and Multiply Aligned TV Series Dataset. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC '14)*, p. 418–425, Reykjavik, Iceland.

SMITH R. (2007). An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR '07)*, p. 629–633, Washington, DC, USA.

WAHLSTER W. (2000). *Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System*, In *Verbmobil: Foundations of Speech-to-Speech Translation*, p. 3–21. Springer Berlin Heidelberg: Berlin, Heidelberg.