Detecting context-dependent sentences in parallel corpora

Rachel Bawden[1] Thomas Lavergne[1] Sophie Rosset[2]
(1) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France
(2) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
lastname@limsi.fr

RÉSUMÉ

Détection dans des corpus parallèles de phrases dépendantes du contexte

Dans cet article, nous proposons plusieurs approches pour l'identification automatique de phrases parallèles qui nécessitent du contexte linguistique extra-phrastique pour être correctement traduites. Notre objectif à long terme est de construire de façon automatique un jeu de test de phrases dépendantes du contexte afin d'évaluer les modèles de traduction automatique conçus pour améliorer la traduction de phénomènes discursifs et contextuels. Nous fournissons une discussion et une critique qui montrent que les approches actuelles ne nous permettent pas d'atteindre notre but et qui suggère que l'évaluation individuelle de phenomènes est probablement la meilleure solution.

ABSTRACT

In this article, we provide several approaches to the automatic identification of parallel sentences that require sentence-external linguistic context to be correctly translated. Our long-term goal is to automatically construct a test set of context-dependent sentences in order to evaluate machine translation models designed to improve the translation of contextual, discursive phenomena. We provide a discussion and critique that show that current approaches do not allow us to achieve our goal, and suggest that for now evaluating individual phenomena is likely the best solution.

MOTS-CLÉS : traduction automatique, contexte, evaluation, discours.

KEYWORDS: machine translation, context, evaluation, discourse.

1 Introduction

Recent work in Machine Translation (MT) has focused on using information beyond the current sentence boundary to aid translation (Libovický & Helcl, 2017; Wang *et al.*, 2017; Jean *et al.*, 2017). The aim of these *contextual* MT systems is to remedy the flaw of traditional MT of translating sentences independently of each other, in particular to improve the translation of discourse phenomena. Despite the progress made in incorporating linguistic context into MT (Bawden *et al.*, 2018), these gains are often not observable using automatic evaluation metrics, such as BLEU (Papineni *et al.*, 2002), and manual analysis of translations is often anecdotal. Whilst strategies such as producing contrastive sentence pairs to be reranked by MT models is a promising strategy for evaluation (Rios Gonzales *et al.*, 2017; Bawden *et al.*, 2018), producing the test sets is often time-consuming and unrepresentative of real data. Morevoer, the distinction is often lacking between examples that need extra-sentential context to be translated and those that do not.

A very useful addition to the test suites available would therefore be a test set of real, attested

examples that require extra-sentential linguistic context to be correctly translated, as this would enable us to evaluate the progress made by contextual MT models specifically on the most difficult examples. Since manually identifying real sentences is very time-consuming, our long-term goal is to automatically construct such a test set. In this paper, we aim to show that designing and implementing a method of automatically detecting these sentences in a parallel corpus remains problematic, as shown by a reflection on what such a method would entail and preliminary experiments using tools currently at our disposal to implement it.

We begin by discussing the types of phenomena we wish to identify (Section 2) and existing work on evaluating discourse phenomena (Section 3). We then define the goals and principles such an identification method would adhere to (Section 4). Finally, in Section 5, we critique two possible approaches to the problem, suggesting theoretical limitations of each approach. Our hope is for this work to provide the basis for discussion on modelling contextual phenomena in a multilingual setting, in a view to automatically identifying context-dependent sentences in the long term.

2 Context-dependent phenomena

In practice, many sentences can be correctly translated in isolation, without surrounding context, which explains why most MT systems today translate sentences independently of each other. However certain phenomena, mostly related to discourse, whose scope, by definition is defined at the discourse rather than the sentence level of discourse, cannot be systematically and correctly translated without extra context. Examples include anaphoric pronoun translation (Hardmeier & Federico, 2010; Guillou, 2016; Loaiciga Sanchez, 2017), lexical disambiguation (Carpuat & Wu, 2007; Rios Gonzales *et al.*, 2017) and cross-lingual discourse connective prediction (Meyer & Popescu-Belis, 2012). These phenomena have a common characteristic: they are cross-lingually ambiguous and can only be disambiguated with the help of linguistic context. This linguistic context can appear within the sentence containing the ambiguous element or elsewhere in the text, in which case we refer to it as extra-sentential linguistic context.

Cross-lingual ambiguity occurs because of mismatches in the language systems of the source and target languages, such that there are several translations possible out of context and only one correct in context.[1] The ambiguity can be morphological, syntactic, semantic and/or discursive. One example of the morphological level is the translation of anaphoric pronouns, which poses difficulties in MT due to structural differences in gender marking cross-lingually. For example, the French translation of English *it* is ambiguous between variants *il* (masc.) and *elle* (fem.), depending on the gender of the French noun with which the pronoun corefers. On the syntactic level, ambiguity can arise from an inherent ambiguity in the source language that is not preserved in the target language. For example, English 'green chestnuts and pears' is ambiguous between French 'des marrons verts et des poires' (only the chestnuts are green) and 'des marrons et des poires verts' (chestnuts and pears are green). Cross-lingual semantic ambiguity is where semantic ambiguity in the source language is not preserved in the target language and a choice must be made between the different meanings. For example, the English word *spade* is ambiguous between French *bêche* 'gardening implement' and *pique* 'suit of cards' (Cf. work on word sense disambiguation in MT by Carpuat & Wu (2007)).[2]

[1] This is different from the choice between synonyms or paraphrases, where any of the choices may be a correct translation.

[2] Although this example may seem contrived, in reality, in spoken dialogue scenarios, where utterances can be very short (E.g. "You've got a spade?"), more ambiguity can be expected.

Finally, at the discursive level, elements such as discourse connectives are often language-specific and are often expressed differently cross-lingually (Cf. work on implicitation of discourse connectives by Meyer & Webber (2013)).

Although the examples given above are relatively well-studied phenomena, particularly in a monolingual setting (coreference resolution, word sense disambiguation, discourse relation labelling, etc.), this cannot be seen as an exhaustive list of context-dependent phenomena. Ideally, it would be useful to study translation quality on all types of context-dependent sentences, not just those in a pre-defined list, especially as they are dependent on the particular language pair.

3 Evaluating discourse in MT

Evaluating discourse and other context-dependent phenomena in MT poses a problem for two main reasons.[3] Firstly, most sentences do not require context to be translated. When they do, few words are affected by an incorrect translation relative to the total number of words in the dataset, despite the fact that these errors can be seriously detrimental to the understanding of the translation.[4] Secondly, the correct translation of certain discourse phenomena, including anaphoric pronoun prediction, depends on previously made translation choices (ensuring translation coherence), ruling out metrics that rely on comparing surface forms of the predicted translation with a reference translation.

As interest in contextual MT surges, the question of how to correctly evaluate the impact of the added context has not been far behind, with different solutions for evaluation, both manual and automatic, being proposed, aimed to overcome the problems described above. In terms of manual evaluation, the aim has been to construct a corpus containing only examples of interest in order to combat the sparsity problem cited above. Isabelle *et al.* (2017) provides such a set of test examples designed to test different well-known problems faced by MT systems, including discourse phenomena. Two solutions have been proposed for automatic evaluation of specific phenomenon. The first, adopted by shared task organisers for both the cross-lingual pronoun prediction task at WMT'16 and DiscoMT'17 (Guillou *et al.*, 2016; Loáiciga *et al.*, 2017) and the cross-lingual word sense disambiguation (WSD) task at SemEval-2013 (Lefever & Hoste, 2013), is to change the nature of the task, and to evaluate the models' ability to translate solely the word of interest, whilst the rest of the translation is imposed for all contestants. This alleviates the second problem of translation coherence. The second automatic method, which also involves avoiding comparison of different models' translations, is to evaluate the capacity of MT models to rerank different hypotheses. Presenting models with contrastive pairs of examples and comparing the models on their ability to rank the correct hypothesis higher than an incorrect one is a way of indirectly evaluating them (Cf. (Sennrich, 2017) for grammatical errors (Rios Gonzales *et al.*, 2017) for WSD and (Bawden *et al.*, 2018) for coreference and lexical coherence/cohesion.

Aside from (Bawden *et al.*, 2018), which imposes that the disambiguating context occur in the previous sentence, the other automatic evaluation methods do not control for the fact that the disambiguating context can appear within the current sentence or beyond the sentence boundary.[5] This means that many examples in the sets can be resolved using sentence-internal context and therefore do

[3] Cf. (Hardmeier, 2012) for a detailed overview of problems faced.

[4] E.g. a coreference error typically leads to one mistranslated pronoun, which changes the entire meaning of the sentence.

[5] The number of pronouns with intra-sentential antecedents was roughly equal to the number with extra-sentential antecedents in the DiscoMT2015 test set (Guillou, 2016, pp. 161).

not directly evaluate the ability of contextual models to use context beyond the current sentence. This notably proved problematic for the evaluation of the 2016 pronoun task, of which the highest performing model did not use any extra-sentential context (achieved higher scores based on the inter-sentential examples alone). A useful complement to these test suites would therefore be a method of automatically constructing a test set of sentences that require linguistic context to be correctly translated. The advantages of such a method would be its automatic nature, given the difficulty of manually finding representative examples of context-dependent phenomena and the fact that it could potentially find more diverse phenomena than a human annotator is capable of finding.

4 Automatic context-dependent sentence detection

Our long-term goal is to propose and develop a method of identifying real corpus examples that are cross-lingually ambiguous and necessarily require extra-sentential context (as opposed to intra-sentential context) to be correctly translated.[6] In theory, such a method would separate parallel sentences for which all information needed to produce the target sentence is found within the source sentence (non-context-dependent) from those for which part of the information can only be found in the surrounding sentences (context-dependent).

4.1 Goals and principles

To achieve our goal, the ideal method would adhere to a certain number of principles to ensure (i) the unbiased nature of the test set, (ii) diversity and a large coverage of the phenomena detected and (iii) easy transferability to other language pairs. Although these properties may not be mutually attainable, attempting to adhere to these three properties is key to developing a detection method.

(i) Unbiased test set A test set should be inherently unbiased towards a certain MT model or a certain type of model if it is to be used to fairly evaluate and compare models. This means that, ideally, the detection method itself should not rely on an existing MT model whose goal is to accomplish a task that the test set is designed to test. In our specific case, this means that any use of contextual MT models would violate this principle.

(ii) Diversity and large coverage of phenomena A number of cases have been previously identifed in the literature as requiring context to be correctly translated, for example anaphoric pronouns, lexical ambiguity, discourse connectives, other cases of lexical cohesion. However, in practice, the main focus has been on only a couple of these phenomena, namely anaphoric pronoun resolution, and to a lesser extent lexical ambiguity. It is therefore interesting to keep the method as generic as possible, giving us the opportunity of identifying new context-dependent phenomena.

(iii) Easy transferability to other language pairs To ensure that similar test sets can be easily produced for other language pairs, the detection method should be independent or at least only weakly dependent on the language pair. Since the majority of contextual phenomena depend on the language systems of the source and target language, this third point complements the previous point concerning the diversity of linguistic phenomena; the less *a priori* knowledge of the language pair required, the more adaptable the method will be to new language pairs, for which we do not have such knowledge.

[6] We make the approximation that we can judge whether a source sentence is cross-lingually ambiguous based on the reference translation in the parallel corpus. In reality a translation can be chosen that avoids the ambiguity altogether.

The question is, is such a method currently possible?

5 Comparison of methods

An ideal method would be one relying on complete and comparable representations of the source sentence and of the target sentence both with and without linguistic context. Intuitively, for context-dependent sentences to be correctly translated, the information present in the representation of the target sentence would be impossible to reconstruct from the representation of the source sentence, unless the information from the context is also included. We look at two different approaches for simulating this idealised scenario, working (i) at the sentence level and (ii) at the word level.

Modelling at the sentence level

Following promising work on distributional representations of words (Mikolov *et al.*, 2013; Pennington *et al.*, 2014), recent work has emerged on the distributional representation of larger units of text, such as sentences. These representations are meant to encode generic, often semantic information about the sentence in fixed-size vectors. If sentence embeddings can encode information about a sentence, can they provide the necessary framework to determine whether or not a target sentence is translatable from its source sentence alone, ignoring its context? A positive answer to this question would require the following to be true: (i) a neural network model can be trained to predict the target sentence embedding from the source sentence embedding; a poor prediction for a given source embedding would be a sign that all the information necessary to produce its corresponding target embedding is not present in the source embedding; (ii) a second model trained to predict the target sentence embedding from a joint embedding of the source sentence and its context (source- or target-side) would predict a better target embedding for this context-dependent sentence.

The problem with this method is the number of assumptions that are made: (i) the sentence embedding fully represents the sentence, (ii) a mapping can be learnt between source and target sentence embeddings, (iii) we have a reliable metric to evaluate whether the contextually predicted sentence embedding is significantly more similar to the real target sentence embedding than the non-contextual one. Preliminary exploratory experiments in this direction which aimed to learn the mapping between DOC2VEC embeddings (Mikolov *et al.*, 2013) in the source and target languages using a small feedforward neural network confirmed that these assumptions were too great. One fundamental flaw with such an approach is that we have little control over the type of information stored in the representation, and no guarantee that this information will be useful for predicting cross-lingual ambiguity. With no control over the type of information modelled, evaluating whether the predicted embedding is sufficiently similar to the true target representation is also an open problem, and makes the method untractable. Given an imperfect representation of a sentence, judging whether a prediction is more similar to the target representation than another is impossible without knowing on what criteria we base the similarity. The approach could only really work with a near-perfect representation of all the information in a sentence, or more control over what kind of information is stored. Given our very generic aim to identify all types of context-dependent phenomena, this approach is not yet feasible. The problem is almost circular; if we had a method to perfectly map the representation of a sentence in context from one language to another, machine translation itself would be a solved task.

Modelling at the word level

Given the problem of obtaining sufficiently complete sentence-level embedding representations, a

reasonable compromise is to try to work at the word level. We therefore consider a second, reduced approach, this time assuming that the ambiguity arises from a single word in the source sentence and only affects its translation in the target sentence.[7] We therefore also need to make the assumption that we have a method to identify sentences containing an ambiguous element. This splits the problem into two steps: (i) identifying sentences containing ambiguous elements, and (ii) separating the sentences which do not need extra context to be translated from those that do. Given that methods exist to detect specific phenomena in corpora, e.g. anaphoric pronouns (Hardmeier *et al.*, 2015) and semantically ambiguous words (Rios Gonzales *et al.*, 2017), we suppose that new methods can be developed for more phenomena. This reduces the task to identifying whether the disambiguating context is found within the sentence, in the neighbouring sentences or cannot be found in the text at all.

An approach at the word level would typically look at the probability of the ambiguous target word given just the current sentence and also given sentence-external context, compared to the probability of the alternative, incorrect solution(s). For example, in the *le chat miaule et [ ] dort* for "the cat meow and **it** sleeps", where the target word is *il* and the incorrect alternative *elle*, we would expect the target word to have a higher probability than the alternative word, regardless of the addition of extra context, since coreference is resolved within the sentence. In a sentence where the antecedent appears in the previous sentence, we would expect the probability of the target word to increase with the addition of this previous sentence relative to the probability of the alternative translation.

Yet again, this method suffers from strong assumptions about the capacity of current NLP models to use contextual information to make such predictions. The assumptions were confirmed to be false by exploratory experiments using tool CONTEXT2VEC (Melamud *et al.*, 2016), which can be used like a language model to make predictions about the form of a target word given a certain context. Three limitations were observed: (i) the intrinsic probability of a word, as determined by its frequency, has a very large effect on its probability in context, making it very complicated to assess the effect of adding context, (ii) the capacity of generic language models to model complex and structured problems such as coreference chains is insufficient, even for simple, short utterances, and (iii) in light of the second limitation, all context, even if not directly relevant to the translation of the ambiguous word, has an effect on the probability of the word. We have little control over which information is considered important by the model, particularly if we wish to keep the approach as general as possible.

6 Conclusion

We have described and motivated a theoretically interesting task of identifying sentences that are cross-lingually ambiguous and dependent on extra-sentential linguistic context. Beyond translation, this could have a wider impact on NLP applications, including dialogue generation and understanding. Through a reflection on the pre-requisites for such a detection method, and by exploring two different approaches to the problem, we have found that the task is very ambitious. The limitations identified have shown us that as long as complete and robust representations of all information within sentences are not yet achievable, the task of identifying context-dependent sentences using a method that is agnostic to the type of phenomenon is unlikely to be attainable. For now it appears that detecting contextual phenomena is better performed on a per-phenomenon basis.

[7] This assumption works at least for most of the cases cited in Section 2.

References

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies*, New Orleans, Louisiana, USA. To appear.

CARPUAT M. & WU D. (2007). Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In *Proceedings of the 11th Machine Translation Summit*, p. 73–80, Copenhagen, Denmark.

GUILLOU L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, School of Informatics. University of Edinburgh.

GUILLOU L., HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y., CETTOLO M., WEBBER B. & POPESCU-BELIS A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 1st Conference on Machine Translation*, WMT'16, p. 525–542, Berlin, Germany.

HARDMEIER C. (2012). Discourse in statistical machine translation. a survey and a case study. *Discours [online]*, **11**.

HARDMEIER C. & FEDERICO M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, IWSLT'10, p. 283–289, Paris, France.

HARDMEIER C., NAKOV P., STYMNE S., TIEDEMANN J., VERSLEY Y. & CETTOLO M. (2015). Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, DISCOMT'15, p. 1–16, Lisbon, Portugal.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP'17, p. 2476–2486, Copenhagen, Denmark.

JEAN S., LAULY S., FIRAT O. & CHO K. (2017). Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. arXiv: 1704.05135.

LEFEVER E. & HOSTE V. (2013). SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, p. 158–166, Atlanta, Georgia.

LIBOVICKÝ J. & HELCL J. (2017). Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, p. 196–202, Vancouver, Canada.

LOAICIGA SANCHEZ S. (2017). *Pronominal anaphora and verbal tenses in machine translation*. PhD thesis, University of Geneva.

LOÁICIGA S., STYMNE S., NAKOV P., HARDMEIER C., TIEDEMANN J., CETTOLO M. & VERSLEY Y. (2017). Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, DISCOMT'17, p. 1–16, Copenhagen, Denmark.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, CoNLL'16, p. 51–61, Berlin, Germany.

MEYER T. & POPESCU-BELIS A. (2012). Machine Translation of Labeled Discourse Connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas*, AMTA'12, p. 129–138, San Diego, California, USA.

MEYER T. & WEBBER B. (2013). Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st Workshop on Discourse in Machine Translation*, DISCOMT'13, p. 19–26, Sofia, Bulgaria.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, NIPS'13, p. 3111–3119, Lake Tahoe, USA.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02, p. 311–318, Philadelphia, Pennsylvania, USA.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, p. 1532–1543, Doha, Qatar.

RIOS GONZALES A., MASCARELL L. & SENNRICH R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the 2nd Conference on Machine Translation*, WMT'17, p. 11–19, Copenhagen, Denmark.

SENNRICH R. (2017). How Grammatical is Character- level Neural Machine Translation? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'17, p. 376–382, Valencia, Spain.

WANG L., TU Z., WAY A. & QUN LIU (2017). Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP'17, p. 2816–2821, Copenhagen, Denmark.