Explorations in Using Grammatical Dependencies for Contextual Phrase Translation Disambiguation

Aurélien Max[1], Rafik Makhloufi[1], and Philippe Langlais[2]

[1] LIMSI-CNRS & Université Paris-Sud 11, Orsay, France
[2] DIRO, Université de Montréal, Canada

Abstract. Recent research has shown the importance of using source context information to disambiguate source phrases in phrase-based Statistical Machine Translation. Although encouraging results have been obtained, those studies mostly focus on translating into a less inflected target language. In this article, we present an attempt at using source context information to translate from English into French. In addition to information extracted from the immediate context of a source phrase, we also exploit grammatical dependencies information. While automatic evaluation does not exhibit a significant difference, a manual evaluation we conducted provides evidence that our context-aware translation engine outperforms its context-insensitive counterpart.

1 Introduction

One notable shortcoming of the now standard phrase-based approach to Statistical Machine Translation (PBSMT) [1] is that a unique conditional probability distribution $p(e|f)$ is considered for translating all the occurrences of a source phrase f,[3] while obviously, differences in translation may happen due to the context in which the phrase occurs.

A systematic inspection of the output of a state-of-the-art PBSMT engine has been conducted by [2]. Among other interesting things, it shows that sense errors (including wrong lexical choice and source ambiguity) account for 21.9% of the observed errors when translating from English into Spanish, and for 28.2% when translating from Spanish into English. Incorrect forms (including incorrect verb tense or person, and incorrect gender or number agreement) account for 33.9% of errors when translating from English into Spanish, and for 9.9% from Spanish into English. Although the latter type of errors may be recovered to some extent by a target language model, those translation errors essentially arise because the engine fails to use grammatical dependencies present in the source sentences, such as subject-verb or verb-object relations.

The integration of source contextual information in phrase-based SMT already gave rise to a number of interesting works [3–7]. [3] and [4] embedded

[3] We use the standard notation where f is a source phrase, and e is a target phrase.

existing word-sense disambiguation systems into a phrase-based SMT system. Both systems use local collocations and word and POS from the immediate context. In [3], bag-of-word context and basic dependency features are used. [5] trained Support Vector Machine classifiers for every possible candidate translation of a source phrase. They considered words of the immediate context (5 tokens to the left and to the right), n-grams (up to size 3) of words, POS and lemmas, as well as chunking labels. [6] trained a global memory-based classifier that performs implicit smoothing of the probability estimates. The classifier makes use of words and POS information of the immediate context of the phrase (2 tokens to the left and to the right).

The experimental conditions and the gains reported in the aforementioned studies differ significantly. However, it is interesting to note that all but one attempts are considering English as the target language. Given the fact that this type of integration only requires linguistic analysis for the source language, we may interpret this as a particular difficulty in improving performances when translating into a highly inflected language. Translating into a morphologically rich language from a highly inflected language (*e.g.* Spanish → French) should intuitively be easier than translating from a less inflected language. The work by [7] is, to our knowledge, the only attempt that deals with the latter. Their experiments conducted on English → German do not show significant improvements when considering context information.

In this paper, we address translating from English into French using features from the immediate context of the source phrases, as well as features from the larger context through the use of grammatical dependencies. To our knowledge, the impact of this latter type of features has not been previously investigated.[4] The paper is organized as follows. We describe our system in section 2. We present experimental results in section 3, and conclude our work and discuss future avenues in section 4.

2 Context-aware PBSMT system

Our context-aware system is directly inspired by the approach of [6]. It consists in the addition to the so-called log-linear combination of features traditionally maximized by a phrased-based engine, context-informed features in the form of conditional probabilities that a target phrase e is the translation of a source phrase f:

$$h_m(f, C(f), e) = \log P(e|f, C(f))$$

where the context $C(f)$ can be any information extracted from the source sentence to translate. Since data sparsity severely impacts the estimation of such probabilities,[5] we followed [6] and used a memory-based approach for estimating the conditional probabilities.

[4] [3] do mention that their classifiers use *basic dependency features*, but their nature is not further described.

[5] Relative frequency would for instance largely overestimate most of the parameters.

In a nutshell, we ask a decision-tree based classifier to produce for the input $\langle f, C(f)\rangle$, a set of weighted class labels representing the possible translations of f. We obtain a posterior probability $P(e|f, C(f))$ by simply normalizing those weights. We used the Tribl hybrid algorithm[6] of the TiMBL software package [8] as a classifier. Building such a classifier is mainly a matter of collecting training examples $\{(\langle f, C(f)\rangle, e)\}$ for all the phrases f seen in context $C(f)$ and translated as e.[7] This classification implicitly performs smoothing by returning the example in the tree matching on most features. In case of an exact match, the actual class (*i.e.* target phrase) seen in the training material is returned. In case of a mismatch, a majority vote is performed.

Approaches such as [5, 6, 4] relied on information extracted from the immediate context of a source phrase. To begin with, we considered this information as well, and represented an example by a fixed-length vector encoding the words of the source phrase f, their POS tags, the POS tags of two words on the left and the two words on the right of f, as well as the associated words. This is illustrated in Figure 1.

`source` our$_1$/PRP [declaration$_2$/NN of$_3$/PRP] rights$_4$/NNs is$_5$/VBZ unique$_6$/JJ
`target` notre$_1$ déclaration$_2$ des$_3$ droits$_4$ est$_5$ unique$_6$
`example` (⟨declaration_of, NN_PRP,`nil`,PRP,NNS,VBZ,nil,our,rights,is⟩,déclaration_des)

Fig. 1. Encoding of the context information of the English phrase *declaration of* aligned with the French phrase *déclaration des*. Numbers show the word alignment. The symbol `nil` is used in place of missing information.

We also took into account information extracted from the dependency parse of the source sentence. More precisely, we considered the dependencies linking tokens of a given phrase to tokens outside this phrase, such as the dependency *poss(declaration,our)* in the above example, which links the inside word *declaration* to the outside one *our*. We selected a set of 16 dependency types (*e.g.*, *poss*-OUT, a possessive dependency out-linking to an outside word) thanks to their Information Gain value we computed on a held-out data set. Each dependency type is represented in the vector by the outside word[8] it involves, or by the symbol `nil`, which indicates that this type of dependency does not occur in the phrase under consideration.

The ordering of the features according to Information Gain values were consistent with that obtained for the Italian → English system of [6]. As expected, the source phrase as well as its concatenated POS tags are the most discriminative features for predicting the translation of the phrase. Immediate context words and POS tags are the next promising features, the right context being

[6] It does perform slightly better than the IGTree algorithm used in [6].
[7] We relied on an in-house version of the standard phrase extraction procedure [1] for collecting the context information required for each source phrase.
[8] Using POS instead of words as dependency targets slightly decreases performance.

more discriminative than the left one. Dependency-based features were found less informative, which can be explained by the fact that often, the immediate context already captures the discriminative information.

In an attempt to boost the selection of the most probable target phrase according to context disambiguation, we added to the log-linear combination a binary feature proposed by [6] which equals 1 for the target phrase that obtains the highest probability $P(e|f, C(f))$, and 0 otherwise.

3 Experiments and evaluation

3.1 Baseline and context-aware systems

We used the French-English Europarl bitext compiled by [9]. To keep the data to a manageable size, we considered a subset of 95,734 sentences that the Stanford parser [10] could parse,[9] for training our global classifier. Following the standard practice, we performed phrase extraction for at most 7 word-long phrases using Giza++ and the `grow-diag-final-and` heuristics [1]. We obtained approximately 11,5M phrases; 3,7M of which are potentially useful for translating the dev and test corpora gathering 475 and 472 bisentences respectively.

We built a baseline system from the set of contextless extracted biphrases. We investigated two context-aware systems as well. System S1 uses the features from the immediate context only and is a replication of the system described in [6]. System S2 uses the same features plus the 16 most informative dependency features found empirically. Following a practical note in [7], we filtered out from the phrase tables of S1 and S2 entries for which $p(e|f) < 0.0002$. This reduces experimentation time dramatically without impacting translation results very much. No such filtering was done for the baseline system.

Models weights were optimized using Minimum Error Rate Training [11], and decoding was performed using the Moses[10] open source PBSMT decoder. All of our translation engines share the same target language model: a Kneser-Ney smoothed trigram model we trained on the French part of the whole Europarl corpus thanks to the SRILM toolkit [12].

For running S1 and S2 systems, we proceeded sentence by sentence. We first classified each sequence of words of a given sentence offline, thus producing a context-aware phrase-table. This table was merged to the *main* one. Since Moses is not designed to handle context-dependent phrases, we had to serialize each token in the source sentence in order to differentiate each repetition of a source phrase in a sentence. We dynamically modified the phrase table accordingly, and applied the reversed operation to the translation produced.

3.2 Evaluation results

Table 1 reports results obtained for automatic and manual evaluation. For the BLEU and NIST metrics, significance testing using paired bootstrap showed no

[9] We used the POS tags output by this parser as well.
[10] http://www.statmt.org/

significant results using 300 samples and $p < 0.05$. As far as BLEU is concerned, S1 is slightly below the baseline, whereas S2 is just above the baseline. Analysis shows that the baseline system tends to use shorter source phrases than the context-aware systems, and in turn S1 uses shorter phrases than S2. These results are coherent with those of [13] which indicate that context-aware systems tend to make more "truly phrasal" lexical choices.

We carried out a manual evaluation to see if differences not shown by automatic metrics would appear. Four native speakers were asked to rank 100 one-best outputs chosen randomly from the test set, and were presented in random order. Each output was thus assigned a rank in $\{1, 2, 3\}$; ties were allowed.[11] Context-aware systems obtained better average mean rank than the baseline system and were preferred more often (an absolute 10% increase). Globally, two judges found S1 and S2 to have similar performance, while the other two preferred S2. Figure 2 shows an example where S2 was found better than S1 and the baseline by all judges.

	automatic		manual			
	BLEU	NIST	avg. rank	%1st	%2nd	%3rd
Baseline	30.89	**6.72**	1.54	64.7	9.3	**26.0**
S1	30.54	6.70	1.39	74.0	12.0	14.0
S2	**31.06**	6.70	**1.34**	**74.7**	**13.3**	12.0

Table 1. Automatic and manual evaluation results.

Src *it should be made clear to all countries that accession to the European Union is not quite ...*
Baseline/S1 *elle doit être clair pour tous les pays que l' adhésion à l' Union européenne n' est pas tout à fait ...*
S2 *il faut préciser clairement à tous les pays que l' adhésion à l' Union européenne n' est pas tout à fait ...*

Fig. 2. Translation outputs produced by the three systems.

4 Discussion and future work

Whereas the need for exploiting source context information in SMT has been clearly identified, results showing improvements in translation quality only consider translation pairs with a less inflected target language. We have presented

[11] Incidentally, S1 (resp. S2) and the baseline produced 11 (resp. 22) identical translations.

experiments in using source context information including grammatical dependency targets while translating from a less inflected to a highly inflected language. While automatic metrics did not show clear gains on translation quality, a manual evaluation we conducted confirms that context-aware systems can produce better translations for highly inflected target languages as well.

We plan to repeat our experiments on more test sets of a bigger size, as well as to increase the size of the corpus used for training the classifier. We then plan to carry out two further experiments: the first one between two highly inflected languages (e.g. French → Spanish) to see if gains are observed in accord with the experiments presented here; the second one between a highly inflected and a less inflected language (e.g. French → English) to see if our approach is competitive with other systems, and, in particular to see if adding grammatical dependency information is beneficial.

Moreover, several classifiers (different methods and/or vector representations) could be used to learn the weights of the individual features of the log-linear combination, as in [7].

References

1. Koehn, P., Och, F.J., , Marcu, D.: Statistical phrase-based translation. In: Proceedings of NAACL/HLT, Edmonton, Canada (2003)
2. Vilar, D., Xu, J., d'Haro, L.F., Ney, H.: Error Analysis of Statistical Machine Translation Output. In: Proceedings of LREC, Genoa, Italy (2006)
3. Carpuat, M., Wu, D.: Context-dependent phrasal translation lexicons for SMT. In: Proceedings of Machine Translation Summit XI, Copenhagen, Denmark (2007)
4. Chan, Y.S., Ng, H.T., Chiang, D.: Word sense disambiguation improves statistical machine translation. In: Proceedings of ACL'07, Prague, Czech Republic (2007)
5. Giménez, J., Màrquez, L.: Context-aware discriminative phrase selection for SMT. In: Proceedings of WMT at ACL, Prague, Czech Republic (2007)
6. Stroppa, N., van den Bosch, A., Way, A.: Exploiting source similarity for smt using context-informed features. In: Proceedings of TMI, Skvde, Sweden (2007)
7. Gimpel, K., Smith, N.A.: Rich source-side context for SMT. In: Proceedings of WMT at ACL, Columbus, USA (2008)
8. Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A.: TiMBL: Tilburg Memory Based Learner, v6.1, Reference Guide. Technical report, ILK 07-xx (2007)
9. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit, Phuket, Thailand (2005)
10. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, Genoa, Italy (2006)
11. Och, F.J.: Minimum error rate training for statistical machine translation. In: Proceedings of ACL, Sapporo, Japan (2003)
12. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of ICSLP, Denver, Colorado (Sept 2002)
13. Carpuat, M., Wu, D.: Evaluation of context-dependent phrasal translation lexicons for SMT. In: Proceedings of LREC, Marrakech, Morroco (2008)