Contrastive Lexical Evaluation of Machine Translation

Aurélien Max[1,2], **Josep Maria Crego**[1], **François Yvon**[1,2]

(1) LIMSI-CNRS, Orsay, France
(2) Université Paris-Sud 11
{amax,jmcrego,yvon}@limsi.fr

Abstract

This paper advocates a complementary measure of translation performance that focuses on the constrastive ability of two or more systems or system versions to adequately translate source words. This is motivated by three main reasons : 1) existing automatic metrics sometimes do not show significant differences that can be revealed by fine-grained focussed human evaluation, 2) these metrics are based on direct comparisons between system hypotheses with the corresponding reference translations, thus ignoring the input words that were actually translated, and 3) as these metrics do not take input hypotheses from several systems at once, fine-grained contrastive evaluation can only be done indirectly. This proposal is illustrated on a multi-source Machine Translation scenario where multiple translations of a source text are available. Significant gains (up to +1.3 BLEU point) are achieved on these experiments, and contrastive lexical evaluation is shown to provide new information that can help to better analyse a system's performance.

1. Introduction

Automatic evaluation metrics play an important role in helping Machine Translation researchers identify promising improvements to their systems as well as better performing approaches on given tasks. Most automatic metrics are fast to compute, and have been shown to have good correlation with human judgements. However, more and more works resort to human judgments, for example to rank system hypotheses (Callison-Burch et al., 2009) or to integrate human knowledge in semi-automatic metrics (Snover et al., 2009). One can however still question how adequate human judgement is as regards MT evaluation: for example, when ranking system hypotheses, it is often difficult to make clear-cut decisions as one hypothesis may be only locally better than the others. Furthermore, harmful errors such as wrong lexical transfer choices are not always detected and are possibly not much penalized, a drawback shared with typical automatic evaluation metrics.

Additionally, in some cases automatic metrics cannot show significant differences that can be revealed by fine-grained focussed human evaluation. The detailed study of (Vilar et al., 2006) proposes a classification of errors made by MT systems and reports results for their phrase-based SMT system. For example, missing words in the target amounts to 26% of errors in Spanish to English translation (7.2% for content words and 18.8% for filler words), and incorrect target words due to incorrect sense amount to 28.2% of errors. The relatively high amount of errors of those types (totalling more than 50% of errors) emphasizes the importance of taking into account how source words got translated in MT evaluation.

One can argue that the fact that metrics sometimes cannot reveal such results can in fact discourage research on fine-grained phenomena: first because automatic metrics may not report positive results and because human evaluation can be too costly, particularly if many system versions are to be compared; and secondly because standard evaluation test sets may not contain enough instances of impacted words (e.g. source homonyms) or phenomena (e.g. pronominal anaphora) to make improvements measurable.[1]

Another striking characteristic of most automatic evaluation metrics lies in the comparison of a system's hypothesis with one or several reference translations, leading to at least two notable consequences. First, because these metrics do not take input hypotheses from several systems at once, fine-grained contrastive evaluation can only be done indirectly. It is indeed well known that absolute scores are highly dependant on the task and language pair, and fine-grained constrastive evaluation can be very helpful to drive the design of models for specific phenomena.[2] Furthermore, computing some similarity between a system's hypotheses and references puts a strong focus on the target side of translation, and does not allow evaluating translation performance from the source words that were actually translated. It is true, however, that good translations are not always literal word-for-word translations. But because statistical MT, in particular, relies heavily on word alignments, it certainly makes sense to restrict possible translation matches for source words to those that are aligned to them in one or several reference translations.

This paper therefore advocates a complementary measure of translation performance that focuses on the constrastive ability of two systems to adequately translate source words. The remainder of the paper is organized as follows: in section 2. we describe our proposal in more details, making comparisons with other works when relevant. Section 3. presents multi-source Machine Translation by hypothesis selection, an approach to MT which is well suited to illustrate our proposal. We describe experiments in this domain and provide results using both traditional automatic metrics and our proposed constrastive lexical evaluation approach. We finally conclude in section 4.

[1] This also possibly explains why many works artificially constrain their evaluation settings, for example by using small training corpora which make improving systems easier.

[2] One possible solution to this would be the design of ad hoc evaluation data sets, as it is done for several NLP tasks such as Word Sense Disambiguation.

2. Contrastive lexical evaluation of MT

As sketched in the previous section, our approach focuses on the contrastive ability of two systems to accurately translate source words. Source words from a test data file must first be aligned with target words, in order to find their word-by-word reference translations according to one or several sentence-by-sentence reference files. This would ideally be done manually, but automatic alignment can be used directly or manually revised.[3] In any case, the same alignments must be used for all subsequent contrastive measures.

The importance of observing the translation of source words may vary. In particular, one could be more interested in content words than in grammatical words as they have a more direct impact on translation adequacy. More generally, we can restrict the morpho-syntactic categories of observed source words, as well as those of potential target words.

Source words may be aligned to several target words, in which case each target token should be individually searched for in the candidate translation. As with existing evaluation metrics, target words from the reference can only be matched once, and flexible matching may be introduced, based on lemmas, synsets (Lavie and Agarwal, 2007) or more generally translational equivalents (Apidianaki, 2009). Contrary to those metrics, our evaluation does not rely on the (independent) comparison of one system's hypotheses with a reference, but indicates for each individual source word which systems (among two or more systems or system versions) correctly translated it according to some reference translation(s). This allows carrying out detailed contrastive analyses at the word level, or at the level of any word class that is deemed appropriate (e.g. part-of-speech, homonymous words, highly ambiguous words relative to the training corpus[4], etc.)

(Carpuat and Wu, 2008) proposed a brief evaluation of their context-aware phrasal lexicons by showing how many times translations for phrases translated by two systems respectively matched and did not match with the reference translation. This setup was rather limited, as it only considered source phrases in common segmentations for the two systems, considered them equally important with respect to the measure, and used an exact match policy which was possibly missing different though acceptable alternative translations.

[3] In our experiments, we reused the alignment procedure of the moses system (Koehn et al., 2007) on the union of the training and test files. Alignments for the words in the test file are thus strongly influenced by alignements from the training portion of the corpus.

[4] It can be useful to consider the inherent complexity of translating a given word: indeed, translating a homonymous words with equally likely meanings is much harder than translating a word with one predominant sense. Therefore, a match for a hard-to-translate word can be rewarded more. It is also possible to report evaluation results for such hard-to-translate words only. For a statistical MT system, the entropy of the distribution of word translations for a given word learnt from a training corpus can be used as a measure of the difficulty to accurately translate that word given the available data.

3. Illustration: multi-source MT by hypothesis selection

3.1. Multi-source MT

When a text has to be translated, it may be the case that several translations of the text in other languages already exist. Exploiting these existing translations as various knowledge sources is refered to as *multi-source translation* (Och and Ney, 2001; Schwartz, 2008). (Schroeder et al., 2009) evaluate various approaches to multi-source translation, and conclude with the superiority of consensus using confusion networks over the translations produced from each individual language. This exploits the common predictions in the target language from various translations. In this work, we consider a simpler approach based on hypothesis selection. In this illustrative scenario, we aim at improving a single SMT system for a given language pair by exploiting source texts available in several languages. We focus here on improving the lexical choices of a *main* system, by reinforcing its hypotheses that are also proposed by systems translating from other languages to the same target language. For example, the translation of the polysemous French word *avocat* is ambiguous into English (*lawyer* or *avocado*) and also has two distinct translations covering the same senses into Spanish (resp. *abogado* and *aguacate*), but the translation from each of the Spanish translations into English is not ambiguous with respect to the English translations obtained from French. The information about which Spanish word corresponds to the French word can thus allow to choose the correct English word, and should certainly be used when available. Translations proposed from auxiliary languages can therefore be considered as clues that can reinforce choices performed by the main system. This is the main intuition behind the system architecture presented on Figure 1.[5]

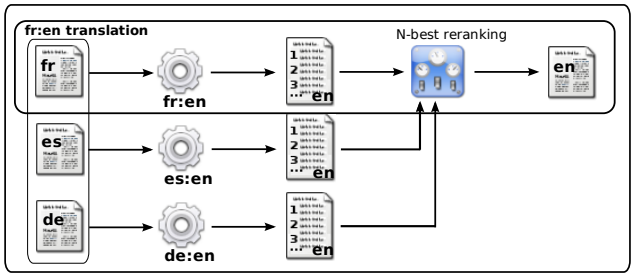

Figure 1: Multi-source SMT by hypothesis selection. The main system translates from French to English, and uses Spanish and German as auxiliary source languages.

Being asymmetrical, our implementation of system combination treats very differently the available sources and systems: the main system will provide the N-best list of hypotheses for reranking, implicitly setting an upper-bound

[5] An important assumption that is made here is that the auxiliary source texts are literal translations of each other that have not undergone deep localization changes. Indeed, important rephrasings, which are sometimes necessary for some language pairs or linguistic phenomena, could prevent the production of sought candidate words in the target language.

over the achievable gains. In comparison, the contribution of auxiliary systems is much weaker, as we are mostly interested in the target words they generate, almost irrespective of their ordering.

3.2. Experiments

We used the Europarl (Koehn, 2005) corpus of parliamentary debates as a source of multilingual parallel text for 11 European languages. In order to study comparable systems, we extracted from the corpus sentences that exist in all languages, using English as pivot. We obtained a smaller corpus of 318,804 lines, corresponding to about 10,3M words for the French part. We used the French→English pair as the main language pair for our experiments.

In this study, we used our own machine translation engine, N-code, which implements the n-gram-based approach to Statistical Machine Translation (Mariño et al., 2006). In a nutshell, the translation model is implemented as a stochastic finite-state transducer trained using a n-gram model of (source,target) pairs (Casacuberta and Vidal, 2004). Training such a model requires to reorder source sentences so as to match the target word order. This is also performed via a stochastic finite-state reordering model, which uses part-of-speech (POS) information[6] to generalize reordering patterns beyond lexical regularities. The reordering model is trained on a version of the parallel corpora where the source sentences have been reordered via the unfold heuristics (Crego and no, 2007), based on a word alignment produced with Giza++[7] run with default settings. The third component of the system, the target language model is a conventional n-gram models of the target language, smoothed with the "improved Kneser-Ney" back-off scheme. Translation takes place in two steps: we first build a source lattice which encodes weighted reordering alternatives of the source sentence; this lattice is then decoded monotonically by searching for the target sentence whose total score maximizes a linear combination of the available scores. The seven coefficients in this linear combination are tuned with an in house implementation of the downhill simplex algorithm (Nelder and Mead, 1965) on development data.

Our baseline system for the main French→English pair is a standard N-code system, which obtained state-of-the-art performance on several evaluations. Baseline system results (BLEU scores) for all language pairs are displayed on Table 1.

In this work, the reinforcement of words that also occur in the translations obtained through auxiliary languages is performed posterior to decoding, in a reranking step. Our aim being to promote those hypotheses from the main system that agree most with auxiliary hypotheses, we compute the n-gram precision of each hypothesis of the main system with respect to each set of auxiliary translations, in a manner analog to the computation of the BLEU metrics (Papineni et al., 2002). For unigram scores, however, only content words[8] are kept, so as to reduce the reward of matching stop-words. For lack of a principled way to generalize this filter to higher-order *n*-grams, no filtering was performed for $n > 1$.

L	fr-L	L-en
da	23.52	29.69
de	18.30	25.77
el	24.15	29.18
es	35.90	31.29
fi	12.68	21.23
it	31.47	28.33
nl	22.87	24.90
pt	33.39	29.73
sv	22.20	31.08

Table 1: BLEU scores for auxiliary systems

For each auxiliary language, we computed the following score, to be interpreted as a cost[9]:

$$score(l) = \sum_{n=1}^{4}(1 - np_n(l)) \quad (1)$$

In Equation (1), np denotes the n-gram precision defined as: $np_n = \frac{|N_n^{hyp} \cap N_n^{ref}|}{|N_n^{hyp}|}$, where N_1^{hyp} and N_1^{ref} correspond respectively to the set of content words in the hypothesis and in the auxiliary hypotheses, and N_n^{hyp} and N_n^{ref} ($2 \geq n \geq 4$) correspond to the set of conventional n-grams. These new scores were then linearly combined, using weights tuned on the development set, with the decoder score; the resulting measure is used to rerank the N-best list of the main system.

In all our experiments, we used a held-out test corpus containing 1000 sentences and 1000-best lists for reranking. Table 2 gives the results obtained using three commonly used automatic metrics (BLEU, TER and WER) for the baseline system and our multi-source system using simultaneously nine auxiliary languages. As can be seen, gains of almost 1.3 BLEU, - 3 TER and - 3.2 WER points are achieved over the baseline system, confirming, according to these metrics, the potential gains that can be expected by exploiting several source texts for translation.[10]

These experiments use all available languages. One might wonder whether comparable performances could be obtained with a smaller number of languages. A positive answer would extend the contexts of use and, in particular, would allow working on translation into several languages

[6] Part-of-speech information is only used for English, French, Spanish and German and is computed with the TreeTagger available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`.

[7] `http://www.fjoch.com/GIZA++.html`

[8] A simple filter removing the most frequent types was used.

[9] Using directly BLEU to score the main hypotheses was not possible, as it produces null scores when a single *n*-gram precision value is null, which is clearly undesirable.

[10] These gains appear however quite limited in the light of the results reported in (Schroeder et al., 2009), suggesting that our approach is less suited to the "pure" multi-source case than sophisticated system combination techniques. The performance reported in (Schroeder et al., 2009) are not directly comparable with ours, as we use almost 3.2 times more training data and a different test set. The relative increase of performance we observe here seems nonetheless in the same ballpark than what they achieve with the MultiLattice combination technique.

	baseline	multi-source
BLEU	30.47	31.76
TER	53.73	50.78
WER	58.08	54.89

Table 2: BLEU scores for French→English obtained with nine auxiliary languages.

by successive translations. It would also help detect those auxiliary languages which tend to reinforce wrong lexical choices and which consequently degrade the results of the main system.

Examining all language combinations requires to build and to optimize systems for the whole set of possible configurations involving auxiliary languages. With nine languages, there are $\sum_{k=1}^{9} C_9^k = 511$ such combinations. We considered a heuristic approach instead, based on a greedy search for the best combination. A straightforward implementation of this idea consists in including auxiliary languages in the order of their relative merits: the set of all combinations involving one auxiliary language is evaluated, then the set of all the combinations involving the best auxiliary language and a second language, and successively all sets involving up to nine languages. Proceeding this way requires to build and optimize a reduced number of $\sum_{i=9}^{1} i = 45$ systems. The top part of Table 3 reports automatic metrics evaluation scores obtained by incrementally adding auxiliary languages in the manner described, using BLEU scores for guiding the search.

3.3. Contrastive lexical evaluation

Although metrics based on comparison on the target side can be informative, an interesting question is how much each newly added auxiliary language can help in selecting the most appropriate translation for a source word (provided it is proposed in the system's *N*-best list).[11] Table 3 additionally shows number of source words that were correctly translated by the baseline French→English system but not by a given multi-source system ('worse' raw), and source words that were correctly translated by that system alone ('better' raw). Overall, the best performance is achieved when adding 6 auxiliary languages (Spanish, Swedish, Danish, Portuguese, Greek and German): in this case, there are 87 more words that were correctly translated by our multi-source system than by the baseline system.[12]

It is first quite informative to see that a comparatively high number of source words which were correctly translated by the baseline system are not correctly translated in all versions of the multi-source systems, although BLEU, for instance, only reports global improvements. The case of the addition of Portuguese provides a clear illustration of what constrative lexical evaluation can reveal: whereas automatic metrics only indicate slight improvements, we can see a significant improvement in the number of words correctly translated by the baseline system and that are not incorrectly translated by the multi-source system anymore (drop from 345 to 306).

Although only a relatively small amount of words have different translations between any two systems in this type of study, it is possible to track down any such local changes. This can be done either at the level of individual word, or at the level of any word class: for example, Figure 2 shows relative improvement for source part-of-speeches corresponding to content words and translated as content words. It is here striking to see that, for instance, the implemented multi-source approach tends to slightly degrade the performance on the translation of adverbs, but improves the translation of verbs, nouns and adjectives. Looking at the case of Portuguese again, its positive impact on the translation of verbs and nouns is confirmed, but we can observe a negative impact on the translation of pronouns.

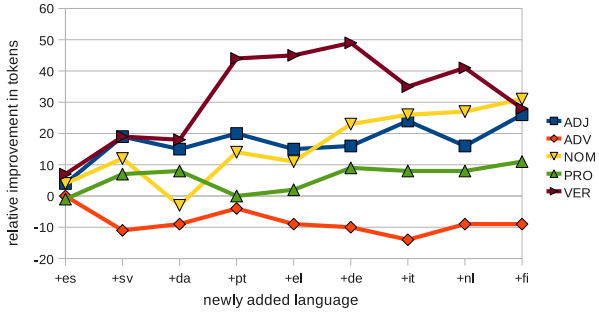

Figure 2: Relative improvement in correctly translated tokens per POS by incrementally adding languages

4. Conclusion

This work proposes a complementary view on the evaluation of MT output which focusses on the contrastive ability of several systems or system versions to accurately translate source words. We illustrated the type of insights that our proposal can provide to drive the design of new translation models on a multi-source statistical MT system, but any type of MT system can be compared with this approach. This work allows tracking fine-grained improvements between systems for frequent error types (Vilar et al., 2006), possibly without recourse to any further human annotation.[13]

Acknowledgements

This work was partly realised as part of the Quaero Program, funded by OSEO, the French agency for innovation.

5. References

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *Proceedings of EACL*, Athens, Greece.

[11] It is to be noted that for a given word in the main source language we cannot expect literal translation to exist in all auxiliary language source text, which explains in part the better performance of a system output combination approach on that task (Schroeder et al., 2009).

[12] Note that this measure can be applied to any pair of systems.

[13] As mentioned previously, using manually annotated word alignments between the source text and the available references would make the evaluation even more precise.

Order of (cumulative) addition of languages to the multi-source system										
Languages	bsl	+es	+sv	+da	+pt	+el	+de	+it	+nl	+fi
Absolute automatic metric scores										
BLEU	30.47	31.05	31.49	31.52	31.54	31.54	31.60	**31.79**	31.78	**31.76**
TER	53.73	52.45	51.51	51.59	51.38	51.40	51.09	50.88	50.87	**50.78**
WER	58.08	56.67	55.69	55.80	55.54	55.53	55.20	55.02	55.00	**54.89**
Contrastive results relative to baseline system (in number of correctly translated source words)										
worse	-	**299**	339	345	306	305	307	319	334	327
better	-	313	385	374	380	369	394	398	**417**	**414**
rel. gain	-	+14	+46	+29	+74	+64	**+87**	+77	**+83**	**+87**

Table 3: Evaluation results obtained when adding auxiliary languages using greedy search

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

Marine Carpuat and Dekai Wu. 2008. Evaluation of context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of LREC 2008*, Marrakech, Morroco.

Francesco Casacuberta and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

Josep M. Crego and Jose B. Mariño. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, Phuket, Thailand.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

José Mariño, Rafael E. Banchs, Josep Maria Crego, Adria de Gispert, Patrick Lambert, J.A.R. Fonollosa, and Martha Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549.

J.A. Nelder and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of MT Summit*, Santiago de Compostela, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, USA.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 719–727, Athens, Greece.

Lane Schwartz. 2008. Multi-source translation methods. In *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 279–288, Waikiki, Hawaï.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL 2009*, Athens, Greece.

David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of LREC*, Genoa, Italy.