Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 103-110 recibido 17-04-14 revisado 07-06-14 aceptado 10-06-14

Document-Level Machine Translation as a Re-translation Process*

Traducción Automática a Nivel de Documento como Proceso de Retraducción

Eva Martínez Garcia
Cristina España-Bonet
TALP Research Center
Univesitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
emartinez@lsi.upc.edu y cristinae@lsi.upc.edu

Lluís Màrquez
Qatar Computing Research Institute
Qatar Foundation
Tornado Tower, Floor 10,
P.O. Box 5825, Doha, Qatar
lluism@lsi.upc.edu

Resumen: Los sistemas de Traducción Automática suelen estar diseñados para traducir un texto oración por oración ignorando la información del discurso y provocando así la aparición de incoherencias en las traducciones. En este artículo se presentan varios sistemas que detectan incoherencias a nivel de documento y proponen nuevas traducciones parciales para mejorar el nivel de cohesión y coherencia global. El estudio se centra en dos casos: palabras con traducciones inconsistentes en un texto y la concordancia de género y número entre palabras. Dado que se trata de fenómenos concretos, los cambios no se ven reflejados en una evaluación automática global pero una evaluación manual muestra mejoras en las traducciones.
Palabras clave: Traducción Automática Estadística, Discurso, Coreferencia, Coherencia

Abstract: Most of the current Machine Translation systems are designed to translate a document sentence by sentence ignoring discourse information and producing incoherencies in the final translations. In this paper we present some document-level-oriented post-processes to improve translations' coherence and consistency. Incoherences are detected and new partial translations are proposed. The work focuses on studying two phenomena: words with inconsistent translations throughout a text and also, gender and number agreement among words. Since we deal with specific phenomena, an automatic evaluation does not reflect significant variations in the translations. However, improvements are observed through a manual evaluation.
Keywords: Statistical Machine Translation, Discourse, Coreference, Coherence

1 Introduction

There are many different Machine Translation (MT) systems available. Differences among systems depend on their usage, linguistic analysis or architecture, but all of them translate documents sentence by sentence. For instance, in rule-based MT systems, rules are defined at sentence level. In data-based MT systems, the translation of a document as a whole make the problem computationally unfeasible. Under this approach, the wide-range context and the discourse information (coreference relations, contextual coherence, etc.) are lost during translation.

Since this is one of the limitations for current MT systems, it is interesting to explore the possibility of improving the quality of the translations at document level. There are several phenomena that confer coherence to final translations that cannot be seen in an intra-sentence scope, for instance, some pronouns or corefered words spanning several sentences, or words that depend on a specific topic and should be translated in the same way through a document.

Following the path of some recent works (Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Xiao et al., 2011; Hardmeier, Nivre, and Tiedemann, 2012), we study some phenomena paying special attention to lexical, semantic and topic cohesion, coreference

* Supported by an FPI grant within the OpenMT2 project (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation (MICINN) and by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC).

ISSN 1135-5948

and agreement. In general, these tasks can be done following two different approaches. On the one hand, discourse information can be integrated inside a decoder, this is, trying to improve translations quality at translation time. On the other hand, the translation can be thought as a two-pass process where the characteristic phenomena are detected in a first step and re-translated afterwards.

In this work we start from a standard phrase-based Statistical Machine Translation system (SMT) from English to Spanish, and design and develop post-process architectures focusing on the phenomena just mentioned. We introduce a method to detect inconsistent translations of the same word through a document and propose possible corrections. We also present an approach to detect gender and number disagreements among coreferred words, which is extended to deal with intra-sentence disagreements.

The paper is organized as follows. We revisit briefly the related work in Section 2. Section 3 contains the description of the resources that we used to design and run the experiments explained in Section 4. Section 5 shows the results of the different translation systems together with a complete manual evaluation of the selected phenomena. Finally, we draw our conclusions and describe some lines of future work in Section 6.

2 Related Work

In the last years approaches to document-level translation have started to emerge. The earliest approaches dealt with pronominal anaphora within an SMT system (Hardmeier and Federico, 2010; Nagard and Koehn, 2010). These authors develop models that, with the help of coreference resolution methods, identify links among words in a text and use them for a better translation of pronouns. The authors in (Gong, Zhang, and Zhou, 2011) approach the problem of topic cohesion by making available the previous translations at decoding time by using a cache system. In this way, one can bias easily the system towards the lexicon already used.

Document-level translation can be also seen as the post-process of an already translated document. In (Xiao et al., 2011), the authors study the translation consistency of a document and re-translate source words that have been translated in different ways within a same document. The aim is to incorporate document contexts into an existing SMT system following 3 steps. First, they identify the ambiguous words; then, they obtain a set of consistent translations for each word according to the distribution of the word over the target document; and finally, generate the new translation taking into account the results of the first two steps.

All of these works are devoted to improve the translation in one particular aspect (anaphora, lexicon, ambiguities) but do not report relevant improvements as measured by an automatic metric, BLEU (Papineni et al., 2002).

Recently, the authors in (Hardmeier, Nivre, and Tiedemann, 2012) presented Docent, an SMT document-level decoder. The decoder is built on top of an open-source phrase-based SMT decoder, Moses (Koehn et al., 2007). The authors present a stochastic local search decoding method for phrase-based SMT systems which allows decoding complete documents. Docent starts from an initial state (translation) given by Moses and this one is improved by the application of a hill climbing strategy to find a (local) maximum of the score function. The score function and some defined change operations are the ones encoding the document level information. The Docent decoder is introduced in (Hardmeier et al., 2013).

3 Experimental Setup

In order to evaluate the performance of a system that deals with document-level phenomena, one needs to consider an adequate setting where the involved phenomena appear.

3.1 Corpora and systems

Most of the parallel corpora used to train SMT systems consist of a collection of parallel sentences without any information about the document structure. An exception is the News Commentary corpus given within the context of the workshops on Statistical Machine Translation[1]. The corpus is build up with news, that is, coherent texts with a consistent topic throughout a document. Besides, one can take advantage from the XML tags of the documents that identify the limits of the document, paragraphs and sentences.

This corpus is still not large enough to train an SMT system, so, for our baseline system we used the Europarl corpus (Koehn, 2005) in its version 7. All the experiments

[1] http://www.statmt.org/wmt14/translation-task.html

are carried out over translations from English to Spanish.The different morphology between these two languages should contribute to obtain troublesome translations which can be tackled with our methodology.

Our baseline system is a Moses decoder trained with the Europarl corpus. For estimating the language model we use SRILM (Stolcke, 2002) and calculate a 5-gram language model using interpolated Kneser-Ney discounting on the target side of the Europarl corpus. Word alignment is done with GIZA++ (Och and Ney, 2003) and both phrase extraction and decoding are done with the Moses package. The optimization of the weights of the model is trained with MERT (Och, 2003) against the BLEU measure on the News Commentary corpus of 2009 (NC-2009, see Table 1 for the concrete figures of the data).

3.2 Data annotation

Besides the aforementioned markup, the 110 documents in the test set have been annotated with several linguistic processors. In particular, we used the Part-of-Speech (PoS) tagger and dependency parser provided by the Freeling library (Padró et al., 2010), the coreference resolutor RelaxCor (Sapena, Padró, and Turmo, 2010) and the named entity recognizer of BIOS (Surdeanu, Turmo, and Comelles, 2005). Whereas the PoS has been annotated in both the source (English) and the target (Spanish) sides of the test set, named entities, dependency trees and coreferences have been annotated in the source side and projected into the target via the translation alignments (see Section 5).

4 Systems Description

4.1 Document-level phenomena

4.1.1 Lexical coherence

The first characteristic we want to improve is the lexical coherence of the translation. Ambiguous words are source words with more than one possible translation with different meanings. Choosing the right translation is,in this case, equivalent to disambiguate the word in its context. Taking the assumption of "one-sense-per-discourse", the translation of a word must be the same through a document. So, our problem is not related to Word Sense Disambiguation, but we want to identify the words in the source document that are translated in different ways in the target. For example, the English word "desk" appearing in the first News can be translated as "ventanilla", "escritorio", "mesa" or "mostrador" according to the translation table of our baseline system, where the different options are not exact synonyms. The aim of our system is to translate "desk' as "mesa" homogeneously throughout the document as explained in Section 4.2.1. This is an example from the 488 instances of words whit inconsistent translations that we found in our corpus using our baseline system.

4.1.2 Coreference and agreement

It is easy to find words that corefer in a text. A word corefers with another if both refer to the same entity. These words must in principle agree in gender and number since they are representing the same concept (person, object, etc.). For instance, if the term "the engineer" appears referring to a girl as it is the case in News 5, the correct translation in Spanish would be "la ingeniera" and not "el ingeniero".

Once we identify and try to fix incoherences in gender or number inside coreference chains, we can take advantage of the analysis and the applied strategies in the coreference problem to correct agreement errors in the intra-sentential scope. This is, in fact, a simpler problem because it is not affected by possible errors given by the coreference resolutor. However, since dependencies among words are shorter, the expressions tend to be translated correctly by standard SMT engines. In our corpus, we only found two instances where the agreement processing could be applied to coreference chains, so most of our analysis finally corresponds to the intra-sentence agreement case.

4.2 Re-translation systems

Given these phenomena we design two main post-processes to detect, re-translate and fix the interesting cases. Figure 1 shows the basic schema of the post-processes.

4.2.1 Lexical coherence

The first step of the post-process that deals with lexical coherence is to identify those words from the source document translated in more than one way. We use the PoS tags to filter out only the nouns, adjectives and main verbs in English. Then, the word alignments given by a first-pass Moses' translation are used to link every candidate token to its translation, so those tokens aligned with more than one different form in the target

	Corpus	News	Sentences	English Tokens	Spanish Tokens
Training	Europarl-v7	–	1,965,734	49,093,806	51,575,748
Development	NC-2009	136	2,525	65,595	68,089
Test	NC-2011	110	3,003	65,829	69,889

Table 1: Figures on the corpora used for training, development and test.

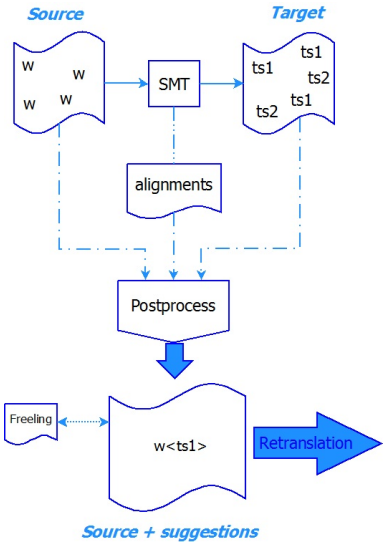

Figure 1: Structure of the post-processes that rewrite the source document in order to prepare it to a retranslation step.

side can be identified. Following the example of Section 4.1.1, if the word "desk" appears three times in the text and it is translated two times as "mesa" and one as "mostrador", then the pair (desk, desk) (mesa, mostrador) will be selected for re-translation.

Re-translation is done in two different ways: *restrictive* and *probabilistic*. The *restrictive* way forces[2] as a possible translation the most used option in the current document; in case of tie, there is no suggestion in order to avoid biasing the result in a wrong way. By doing this, we somehow control the noise introduced by the post-process but we also lose information given by the decoder in the available translation. On the other hand, the *probabilistic* way suggests the most used options as possible translations assigning them a probability estimated by frequency counts within the options. So, in this case, one feeds the decoder with the most suitable options and let it choose among them. This option introduces more noise because the system is managing more possible translations than in the previous situation, sometimes as many as in the initial state translation.

[2] *Forcing* or *suggesting* a translation is a feature available in the Moses decoder that involves an XML markup of the source sentence with the information of translation options.

4.2.2 Gender and number agreement

The post-process for disagreements in gender and number analyses every source document and annotates it with PoS, coreference chains, and dependency trees. The main structure in this case is the tree since it is the one that allows to link the elements that need to agree. A tree traversal is performed in order to detect nouns in the source. When a noun is found and its children are determiners and/or adjectives, the matching subtree is projected into the target via the SMT word alignments. In the target side, one can check the agreement among tokens by using the PoS tags. If there is a disagreement, the correct tag for the adjective or determiner is built using the corresponding Freeling library, which allows to get the correct form in the target language for the translation.

The system implements a similar strategy to check the agreement among subject and verb. A tree traversal allows to detect the node that represents the verb of the sentence and the child corresponding to the subject. The structure is projected into the target via the alignments and the agreement is verified using the PoS information. If the subject is a noun, we assume that the verb must be conjugated in third person plural or singular depending on the number of the noun; if it is a pronoun, gender, person and number must agree. As before, if there is a disagreement, the system assigns the correct tag to the verb, and the form is generated using the Freeling library.

In both cases (determiner–adjective(s)–noun(s) and subject–verb) the output of the pre-process is a proposed new translation for translations that show a disagreement. Similarly to the *restrictive* and *probabilistic* systems of the previous subsections, here we run the re-translation step in two ways: forcing an output with new translation or allowing the interaction of this new translation

System	BLEU	NIST	TER	METEOR	ROUGE	SP-Op	ULC
Baseline	26.73	7.34	55.45	27.78	29.36	31.53	85.01
Lex_R	26.76	7.34	55.39	27.80	29.39	31.60	83.26
Lex_P	26.73	7.34	55.41	27.77	29.38	31.58	85.07
Agr_R	26.66	7.33	55.46	27.75	29.41	31.69	85.10
Agr_P	26.73	7.33	55.45	27.75	29.41	31.64	85.05
Seq_R_R	26.65	7.32	55.46	27.74	29.40	31.68	85.08
Seq_R_P	26.73	7.33	55.45	27.75	29.40	31.63	85.05
Seq_P_R	26.64	7.32	55.48	27.74	29.38	31.67	79.28
Seq_P_P	26.72	7.32	55.46	27.74	29.40	31.63	85.04

Table 2: Automatic evaluation of the systems. See text for the system and metrics definition.

with the remaining translation options of the phrase table. In both cases, the full sentence can be re-translated to accommodate the new options.

The following section shows an automatic and manual evaluation of these systems for the English–Spanish language pair.

5 Experimental Results

The most straightforward way to evaluate translation engines is using automatic metrics on the full test sets. However, in our case, the measures are not informative enough considering that we apply small modifications to previous translations. As an example, Table 2 shows the automatic evaluation obtained with the Asiya toolkit (González, Giménez, and Màrquez, 2012) for several lexical metrics (BLEU, NIST, TER, METEOR and ROUGE), a syntactic metric based on the overlap of PoS elements (SP-Op), and an average of a set of 27 lexical and syntactic metrics (ULC).

The first row shows the results for the baseline system built with Moses without any re-translation step (*Baseline*). The second block includes the experiments on lexical coherence alone both restrictive and probabilistic (*Lex_R* and *Lex_P*) and the third block the experiments on agreement alone also in the two cases (*Agr_R* and *Agr_P*). Finally, the last block shows the result of the sequential process with the four combination of systems (*Seq_R_R*, *Seq_R_P*, *Seq_P_R*, *Seq_P_P*). As it can be seen, the scores do not show any systematic preference for a system and it is necessary a manual evaluation of the outputs to study the performance of the re-translation systems.

System	BLEU	tags	words	OK/ch	linTags	linDif
News20bl	13.40					
News20_R	13.56	26	8	5/9	13	6
News20_P	13.22	45	15	7/11	19	8
News25bl	14.42					
News25_R	14.45	18	4	4/4	16	3
News25_P	14.52	38	10	5/5	28	7
News39bl	28.49					
News39_R	28.20	16	5	5/5	15	4
News39_P	28.56	34	11	6/8	25	7
News48bl	30.05					
News48_R	30.06	42	3	3/3	23	10
News48_P	29.83	53	7	4/5	24	15
News49bl	25.54					
News49_R	25.87	24	5	5/5	17	8
News49_P	25.83	42	12	7/8	23	10

Table 3: Manual evaluation of the system for lexical coherence (*Lex* in Table 2) for a subset of news with restrictive and probabilistic systems. See text for column's meaning.

5.1 Manual evaluation

In order to manually evaluate the output of the previous systems, we chose those documents where we include more suggestions into the re-translation step.

In Table 3, one can see the results of evaluating the system devoted to improve the global lexical coherence for the five news where the post-process introduces more changes. For every document, the *News*bl* row represents the scores for the translations obtained using the baseline system. Column *tags* shows the number of introduced tags, *words* the number of different words involved in the tags, *OK/ch* shows the number of changes made with respect to the first translation and how many are correct attending to our criteria of having one-sense-per-discourse and the word appearing in the reference translation. Note that the tags in the probabilistic approach (_P) include the ones of the restrictive approach (_R) since the first strategy allows us to suggest possible new translations of a word in more cases,

not only when we find the strictly most used translation option for a word. In order to see the scope of the introduced changes, *linTags* shows the number of tagged lines in the source text and *linDif* shows the number of different lines between the final translation and the translation the system uses at the beginning. In general, in all our experiments we could see very local changes due to the retranslation step that affected mostly the tagged words without changing the main structure of the target sentence nor the surrounding ones.

We observe that as with the full automatic evaluation, the BLEU scores of our experiments differ in a non-significant way from the baseline and this is because we are introducing only a few changes in a document. For instance, when we re-translate News20, the one that makes the largest number of changes, we only change 9 words using the *restrictive* approach and 11 using the *probabilistic* one. In this concrete document the accuracy of our changes is above the 50%, but in general, the restrictive approach obtains a high performance and, in the rest of the documents evaluated (News25, News39, News48 and News49), the accuracy in the changes is of a 100%. The probabilistic approach shows a slightly lower performance with accuracies around 80%.

A clear example of how our system works can be observed in a document that talks about a judgement. The document contains the phrase "the trial coverage" translated in first place as "la cobertura de prueba" where the baseline system is translating wrongly the word "trial". But, our post-process sees this word translated more times through the document as "juicio", identifies it as an ambiguous word and tags it with the good translation form "juicio". But not all the changes are positive as we have seen. For example, in a document appears the word "building" five times, being translated three times as "construcción" and two times as "edificio". For our system, the first option is better as long as it appears more times in the translation than the second one. So, we suggested the decoder to use "construcción" when translates "building" in the document. Doing that, we produce two changes in the final translation that generate two errors with respect to the reference translation although both translation options are synonyms. So, in this case our system moves away the translation from the reference although both translations should be correct.

system	BLEU	OK/ch	dets	adjs	verbs
News5bl	13.74				
News5_R	14.06	23/26	17/19	6/7	0/0
News5_P	13.79	15/26	12/19	3/7	0/0
News6bl	11.06				
News6_R	11.22	19/23	8/11	11/11	0/1
News6_P	11.10	10/23	4/11	6/11	0/1
News22bl	16.23				
News22_R	14.74	17/25	4/8	13/17	0/0
News22_P	14.89	10/25	2/8	8/17	0/0
News27bl	13.15				
News27_R	12.35	22/28	14/19	7/8	1/1
News27_P	12.76	21/28	14/19	7/8	0/1
News33bl	15.09				
News33_R	16.05	18/22	14/16	3/3	1/3
News33_P	15.97	11/22	7/16	2/3	2/3

Table 4: Manual evaluation of the system that deals with the agreement (*Agr* in Table 2) for a subset of news with restrictive and probabilistic systems. See text for column's meaning.

Regarding to the errors introduced by the systems, we find that they are caused mainly by bad alignments which provoke an erroneous projection of the structures annotated on the source, errors in the PoS tagging, untranslated words, or, finally, a consequence of the fact that the most frequent translation for a given word in the initial state is wrong.

If we move on now to the agreement experiment, we observe the results from the manual evaluation of checking number and gender agreement in Table 4. Column *OK/ch* shows the number of introduced changes (correct/total), the *dets* column shows the changes over determiners, *adjs* over adjectives and *verbs* over verb forms.

In this set of experiments, we observe that the changes induced by our post-process have an impact in the BLEU score of the final translation because in this case the number of changes is higher. For instance, in News22, we observe a drop of almost two points in the BLEU score after applying the post-process although many of the changes made after the re-translation are correct. We observe the same behaviour in News27, although in the rest of news is shown an opposite trend. According to the manual evaluation, the restrictive system is better than the probabilistic one and reaches accuracies above 80% in the selected news.

A positive example of the performance of the system is the re-translation of the source

system	BLEU	OK/ch	dets	adjs	verbs
News20bl	13.40				
News20_R_R	13.38	17/19	14/15	3/3	0/1
News20_R_P	13.44	14/19	11/15	2/3	1/1
News20_P_R	13.21	16/17	13/14	3/3	0/0
News20_P_P	13.44	12/17	10/14	2/3	0/0
News25bl	14.42				
News25_R_R	14.68	12/19	9/13	3/6	0/0
News25_R_P	15.09	15/19	10/13	5/6	0/0
News25_P_R	14.39	10/17	6/11	4/6	0/0
News25_P_P	14.82	13/17	8/11	5/6	0/0
News39bl	28.49				
News39_R_R	30.02	20/22	14/16	6/6	0/0
News39_R_P	29.59	18/22	13/16	5/6	0/0
News39_P_R	29.94	19/21	14/16	5/5	0/0
News39_P_P	29.59	17/21	13/16	4/5	0/0
News48bl	30.05				
News48_R_R	29.57	6/6	5/5	1/1	0/0
News48_R_P	29.60	4/6	4/5	0/1	0/0
News48_P_R	29.57	6/6	5/5	1/1	0/0
News48_P_P	29.60	4/6	4/5	0/1	0/0
News49bl	25.54				
News49_R_R	25.82	9/11	3/4	6/7	0/0
News49_R_P	26.02	9/11	3/4	6/7	0/0
News49_P_R	25.63	8/11	3/4	5/6	0/1
News49_P_P	26.02	9/11	3/4	5/6	1/1

Table 5: Manual evaluation of the translation after combining sequentially both post-processes, first applying the disambiguation post-process and, afterwards, checking for the agreement. The notation is the same as in previous tables.

phrase "the amicable meetings". This phrase is translated by the baseline as "el amistosa reuniones", where one can find disagreements of gender and number among the determiner, the adjective and the noun. The system detects these disagreements and after tagging the source with the correct forms and re-translating, one obtains the correct final translation "las reuniones amistosas", where we observe also that the decoder has reordered the sentence.

Regarding to the errors introduced by the system, we observe again that many of them are caused by wrong analysis. For instance, in the sentence "all (the) war cries" which should be translated as "todos los gritos de guerra", the dependence tree shows that the determiner depends on the noun "war" and not on "cries", so, according to this relation, our method identifies that the determiner and the translation do not agree and produces the wrong translation "todos (la) guerra gritos".

These results also show that for our approach it is easier to detect and fix disagreements among determiners or adjectives and nouns than among subjects and their related verbs. In general, this is because our current system does not take into account subordinated sentences, agent subjects and other complex grammatical structures, and therefore the number of detected cases is smaller than for the determiner–adjective–noun cases. Further work can be done here to extend this post-process in order to identify disagreements among noun phrases and other structures in the sentence that appear after the verb.

In order to complete this set of experiments, we run sequentially both systems. Table 5 shows the results for the combination of systems in the same format as in the previous experiment. Once again, we observe only slight variations in BLEU scores but, manually, we see that when the systems introduce changes, they are able to fix more translations than the ones they damage. Also as before, it is easier to detect and fix disagreements among determiners, adjectives and nouns than those regarding verbs because of the same reason as in the independent system.

6 Conclusions and Future Work

This work presents a methodology to include document-level information within a translation system. The method performs a two-pass translation. In the first one, incorrect translations according to predefined criteria are detected and new translations are suggested. The re-translation step uses this information to promote the correct translations in the final output.

A common post-process is applied to deal with lexical coherence at document level and intra- and inter-sentence agreement. The source documents are annotated with linguistic processors and the interesting structures are projected on the translation where inconsistencies can be uncovered. In order to handle lexical coherence, we developed a post-process that identifies words translated with different meanings through the same document. For treating disagreements, we developed a post-process that looks for inconsistencies in gender, number and person within the structures determiner–adjective(s)–noun(s) and subject–verb.

Because we are treating concrete phenomena, an automatic evaluation of our systems does not give us enough information to assess the performance of the systems. A detailed manual evaluation of both systems

shows that we only introduce local changes. The lexical-coherence-oriented post-process induces mostly correct translation's changes when using our restrictive system, improving the final coherence of the translation. On the other hand, for the post-process focused on the analysis of the number and gender agreement, it achieves more than 80% of accuracy over the introduced changes in the manually-evaluated news documents. We also observed that some of the negative changes are consequence of bad word alignments which introduce noise when proposing new translations.

A natural continuation of this work is to complete the post-processes by including in the study new document-level phenomena like discourse markers or translation of pronouns. On the other hand, we aim to refine the methods of suggestion of new possible translations and to detect bad word alignments. As a future work, we plan to introduce the analysis of these kind of document-level phenomena at translation time, using a document-level oriented decoder like Docent.

References

Gong, Z., M. Zhang, and G. Zhou. 2011. Cache-based document-level statistical machine translation. In *Proc. of the 2011 Conference on Empirical Methods in NLP*, pages 909–919, UK.

González, M., J. Giménez, and L. Màrquez. 2012. A graphical interface for MT evaluation and error analysis. In *Proc. of the 50th ACL Conference, System Demonstrations*, pages 139–144, Korea.

Hardmeier, C. and M. Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proc. of the 7th International Workshop on Spoken Language Translation*, pages 283–289, France.

Hardmeier, C., J. Nivre, and J. Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proc. of the Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 1179–1190, Korea.

Hardmeier, C., S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proc. of the 51st ACL Conference*, pages 193–198, Bulgaria.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proc.: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th ACL Conference*, pages 177–180, Czech Republic.

Nagard, R. Le and P. Koehn. 2010. Aiding pronouns translation with co-reference resolution. In *Proc. of Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Sweden.

Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the ACL Conference*.

Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Padró, L., S. Reese, E. Agirre, and A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, India. Global Wordnet Conference.

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL Conference*, pages 311–318.

Sapena, E., L. Padró, and J. Turmo. 2010. A global relaxation labeling approach to coreference resolution. In *Proceedings of 23rd COLING*, China.

Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*.

Surdeanu, M., J. Turmo, and E. Comelles. 2005. Named entity recognition from spontaneous open-domain speech. In *Proc. of the 9th Interspeech*.

Xiao, T., J. Zhu, S. Yao, and H. Zhang. 2011. Document-level consistency verification in machine translation. In *Proc. of Machine Translation Summit XIII*, pages 131–138, China.