Automatic Projection of Semantic Structures: an Application to Pairwise Translation Ranking

Daniele Pighin **Lluís Màrquez**
TALP Research Center
Universitat Politècnica de Catalunya
{pighin,lluism}@lsi.upc.edu

Abstract

We present a model for the inclusion of semantic role annotations in the framework of confidence estimation for machine translation. The model has several interesting properties, most notably: 1) it only requires a linguistic processor on the (generally well-formed) source side of the translation; 2) it does not directly rely on properties of the translation model (hence, it can be applied beyond phrase-based systems). These features make it potentially appealing for system ranking, translation re-ranking and user feedback evaluation. Preliminary experiments in pairwise hypothesis ranking on five confidence estimation benchmarks show that the model has the potential to capture salient aspects of translation quality.

1 Introduction

The ability to automatically assess the quality of translation hypotheses is a key requirement towards the development of accurate and dependable translation models. While it is largely agreed that proper transfer of predicate-argument structures from source to target is a very strong indicator of translation quality, especially in relation to adequacy (Lo and Wu, 2010a; 2010b), the incorporation of this kind of information in the Statistical Machine Translation (SMT) evaluation pipeline is still limited to few and isolated cases, e.g., (Giménez and Màrquez, 2010).

In this paper, we propose a general model for the incorporation of predicate-level semantic annotations in the framework of Confidence Estimation (CE) for machine translation, with a specific focus on the sub-problem of pairwise hypothesis ranking. The model is based on the following underlying assumption: by observing how automatic alignments project semantic annotations from source to target in a parallel corpus, it is possible to isolate features that are characteristic of good translations, such as movements of specific arguments for some classes of predicates. The presence (or absence) of these features in automatic translations can then be used as an indicator of their quality. It is important to stress that we are *not* claiming that the projections preserve the meaning of the original annotation. Still, it should be possible to observe regularities that can be helpful to rank alternative translation hypotheses.

The general workflow (which can easily be extended to cope with different annotation layers, such as sequences of meaningful phrase boundaries, named entities or sequences of chunks or POS tags) is exemplified in Figure 1. During training (on the left), the system receives a parallel corpus of source sentences and the corresponding reference translations. Source sentences are annotated with a linguistic processor. The annotations are projected using training alignments, obtaining *gold* projections that we can use to learn a model that captures correct annotation movements, i.e., observed in reference translations. At test time, we want to assess the quality of a translation hypothesis given a source sentence. As shown on the right side of Figure 1, the first part of the process is the same as during training: the source sentence is annotated, and the annotation is projected onto the translation hypothesis via automatic alignments. The model is then used

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 1–9,
ACL HLT 2011, Portland, Oregon, USA, June 2011.

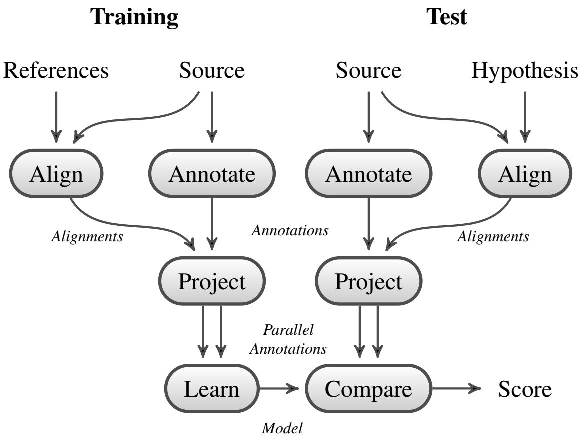

Figure 1: Architectural overview.

to compare the observed projection against the *expected* projection given the source annotation. The distance between the two projections (observed and expected) can then be used as a measure of the quality of the hypothesis.

As it only considers one-sided annotations, our framework does not require the availability of comparable linguistic processors and linguistic annotations, tagsets, etc., on both sides of the translation process. In this way, it overcomes one of the main obstacles to the adoption of linguistic analysis for MT confidence estimation. Furthermore, the fact that source data is generally well-formed lowers the requirements on the linguistic processor in terms of robustness to noisy data, making it possible to employ a wider range of linguistic processors.

Within this framework, in this paper we describe our attempt to bridge Semantic Role Labeling (SRL) and CE by modeling proposition-level semantics for pairwise translation ranking. The extent to which this kind of annotations are transferred from source to target has indeed a very high correlation with respect to human quality assessments (Lo and Wu, 2010a; 2010b). The measure that we propose is then an ideal addition to already established CE measures, e.g., (Specia et al., 2009; Blatz et al., 2004), as it attempts to explicitly model the adequacy of translation hypotheses as a function of predicate-argument structure coverage. While we are aware of the fact that the current definition of the model can be improved in many different ways, our preliminary investigation, on five English to Spanish translation benchmarks, shows promising accuracy on the difficult task of pairwise translation ranking, even for translations with very few distinguishing features.

To capture different aspects of the projection of SRL annotations we employ two instances of the abstract architecture shown in Figure 1. The first works at the *proposition level*, and models the correct movement of arguments from source to target. The second works at the *argument level*, and models the fluency and adequacy of individual arguments within each predicate-argument structure. The *models* that we learn during training are simple phrase-based translation models working on different kinds of sequences, i.e., role labels in the former case and words in the latter. To evaluate the adequacy of an automatically projected proposition or argument, we force the corresponding translation model to generate it (via constrained decoding). The reachability and confidence of each translation are features that we exploit to compare alternative translations, by combining them in a simple voting scheme.

To score systems which are not under our direct control (the typical scenario in CE benchmarks), we introduce a component that generates source-target alignments for any pair of aligned test sentences. This addition has the nice property of allowing us to handle the translation as a black-box, decoupling the evaluation from a specific system and, in theory, allowing the model to cope with phrase-based, rule-based or hierarchical systems alike, as well as with human-generated translations.

The rest of the paper is structured as follows: in Section 2 we will review a selection of related work; in Section 3 we will detail our approach; in Section 4 we will present the results of our evaluation; finally, in Section 5 we will draw our conclusions.

2 Related work

Confidence estimation is the sub-problem within MT evaluation concerned with the assessment of translation quality in the absence of reference translations. A relevant initial work on this topic is the survey by Blatz et al. (2004), in which the authors define a rich set of features based on source data, translation hypotheses, n-best lists and model characteristics to classify translations as “good” or “bad”. In their observations, they conclude that the most relevant features are those based on source/target pairs and on characteristics of the translation model.

Specia et al. (2009) build on top these results by designing a feature-selection framework for confidence estimation. Translations are considered as black-boxs (i.e., no system or model-dependent features are employed), and novel features based on the number of content words, a POS language model on the target side, punctuation and number matchers in source and target translations and the percentage of uni-grams are introduced. Features are selected via Partial Least Squares (PLS) regression (Wold et al., 1984). Inductive Confidence Machines (Papadopoulos et al., 2002) are used to estimate an optimal threshold to distinguish between "good" and "bad" translations. Even though the authors show that a small set of shallow features and some supervision can produce good results on a specific benchmark, we are convinced that more linguistic features are needed for these methods to perform better across a wider spectrum of domains and applications.

Concerning the usage of SRL for SMT, Wu and Fung (2009) reported a first successful application of semantic role labels to improve translation quality. They note that improvements in translation quality are not reflected by traditional MT evaluation metrics (Doddington, 2002; Papineni et al., 2002) based on n-gram overlaps. To further investigate the topic, Lo and Wu (2010a; 2010b) involved human annotators to demonstrate that the quality of semantic role projection on translated sentences is very highly correlated with human assessments.

Giménez and Màrquez (2010) describe a framework for MT evaluation and meta-evaluation combining a rich set of n-gram-based and linguistic metrics, including several variants of a metric based on SRL. Automatic and reference translations are annotated independently, and the lexical overlap between corresponding arguments is employed as an indicator of translation quality. The authors show that syntactic and semantic information can achieve higher reliability in system ranking than purely lexical measures.

Our original contribution lies in the attempt to exploit SRL for assessing translation quality in a CE scenario, i.e., in the absence of reference translations. By accounting for whole predicate-argument sequences as well as individual arguments, our model has the potential to capture aspects which relate both to the adequacy and to the fluency of a translation. Furthermore, we outline a general framework for the inclusion of linguistic processors in CE that has the advantage of requiring resources and software tools only on the source side of the translation, where well-formed input can reasonably be expected.

3 Model

The task of semantic role labeling (SRL) consists in recognizing and automatically annotating semantic relations between a *predicate* word (not necessarily a verb) and its *arguments* in natural language texts. The resulting predicate-argument structures are commonly referred to as *propositions*, even though we will also use the more general term *annotations*.

In PropBank (Palmer et al., 2005) style annotations, which our model is based on, predicates are generally verbs and roles are divided into two classes: core roles (labeled A0, A1, . . . A5), whose semantic value is defined by the predicate syntactic frame, and adjunct roles (labeled AM-*, e.g., AM-TMP or AM-LOC) [1] which are a closed set of verb-independent semantic labels accounting for predicate aspects such as temporal, locative, manner or purpose. For instance, in the sentence "The commission met to discuss the problem" we can identify two predicates, *met* and *discuss*. The corresponding annotations are "[$_{\text{A0}}$ The commission] [$_{\text{pred}}$ met] [$_{\text{AM-PRP}}$ to discuss the problem]" and "[$_{\text{A0}}$ The commission] met to [$_{\text{pred}}$ discuss] [$_{\text{A1}}$ the problem]". Here, A0 and A1 play the role of prototypical subject and object, respectively, and AM-PRP is an adjunct modifier expressing a notion of purpose.

Sentence annotations are inherently non-sequential, as shown by the previous example in which the predicate and one of the arguments of the second proposition (i.e., *discuss* and A1) are completely embedded within an argument of the first proposition (i.e., AM-PRP). Following a widely adopted simplification, the annotations in a sentence are modeled independently. Furthermore we de-

[1] The actual role labels are in the form Arg0, . . . Arg1 and ArgM-*, but we prefer to adopt their shorter form.

scribe each annotation at two levels: a *proposition level*, where we model the movement of arguments from source to target; and an *argument level*, were we model the adequacy and fluency of individual argument translations. The comparison of two alternative translations takes into account all these factors but it models each of them independently, i.e., we consider how properly each propositions is rendered in each hypothesis, and how properly each argument is translated within each proposition.

3.1 Annotation and argument projection

At the proposition level, we simply represent the sequence of role-label in each proposition, ignoring their lexical content with the exception of the predicate word. Considering the previous example, the sentence would then be represented by the two sequences "A0 met AM-PRP" and "A0 * discuss A1". In the latter case, the special character "*" marks a "gap" between A0 and the predicate word. The annotation is projected onto the translation via direct word alignments obtained through a constrained machine translation process (i.e., we force the decoder to generate the desired translation). Eventual discontinuities in the projection of an argument are modeled as gaps. If two arguments insist on a shared subset of words, then their labels are combined. If the projection of an argument is a subset of the projection of the predicate word, then the argument is discarded. If the overlap is partial, then the non-overlapping part of the projection is represented.

If a word insertion occurs next to an argument or the predicate, then we include it in the final sequence. This decision is motivated by the consideration that insertions at the boundary of an argument may be a clue of different syntactic realizations of the same predicate across the two languages (Levin, 1993). For example, the English construct "*A0 give A2 A1*" could be rendered as "*doy A1 a A2*" in Spanish. Here, the insertion of the preposition "*a*" at decoding can be an important indicator of translation quality.

This level of detail is insufficient to model some important features of predicate-argument structures, such as inter-argument semantic or syntactic dependencies, but it is sufficient to capture a variety of interesting linguistic phenomena. For instance, A0-predicate inversion translating SVO into VSO languages, or the convergence of multiple source arguments into a single target argument when translating into a morphologically richer language. We should also stress again that we are not claiming that the structures that we observe on the target side are linguistically motivated, but only that they contain relevant clues to assess quality aspects of translation.

As for the representation of individual arguments, we simply represent their surface form, i.e., the sequence of words spanning each argument. So, for example, the argument representations extracted from "[$_{\text{A0}}$ The commission] [$_{\text{pred}}$ met] [$_{\text{AM-PRP}}$ to discuss the problem]" would be "*The commission*", "*met*", "*to discuss the problem*". To project each argument we align all its words with the target side. The leftmost and the rightmost aligned words define the boundaries of the argument in the target sentence. All the words in between (including eventual gaps) are considered as part of the projection of the argument. This approach is consistent with PropBank style annotations, in which arguments are contiguous word sequences, and it allows us to employ a standard translation model to evaluate the fluency of the argument projection. The rationale here is that we rely on proposition level annotations to convey the semantic structure of the sentence, while at the argument level we are more interested in evaluating the lexical appropriateness of their realization.

The projection of a proposition and its arguments for an example sentence is shown in Figure 2. Here, s is the original sentence and h_1 and h_2 are two translation hypotheses. The figure shows how the whole proposition (p) and the predicate word ($pred$) along with its arguments (*A0*, *A1* and *A2*) are represented after projection on the two hypotheses. As we can observe, in both cases *thank* (the predicate word) gets aligned with the word *gracias*. For h_1, the decoder aligns *I* (A0) to *doy*, leaving a gap between A0 and the predicate word. The gap gets filled by generating the word *las*. Since the gap is adjacent to at least one argument, *las* is included in the representation of p for h_1. In h_2, the projection of A0 exactly overlaps the projection of the predicate ("Gracias"), and therefore A0 is not included in n for h_2.

3.2 Comparing hypotheses

At test time, we want to use our model to compare translation pairs and recognize the most reli-

s	I thank the commissioner for the detailed reply		
h_1	Doy las gracias al comisario por la detallada respuesta		
h_2	Gracias , al señor comisario por para el respuesta		
p	A0 thank A1 A2	**pred**	thank
h_1	A0 +las gracias A1 A2	h_1	gracias
h_2	Gracias A1 A2	h_2	Gracias
A1	the commissioner	**A0**	I
h_1	al comisario	h_1	doy
h_2	al señor comisario	h_2	Gracias
A2	for the detailed reply		
h_2	por la detallada respuesta		
h_2	para el respuesta		

Figure 2: Comparison between two alternative translations h_1 and h_2 for the source sentence s.

able. Let s be the source sentence, and h_1 and h_2 be two translation hypotheses. For each proposition p in s, we assign a confidence value to its representation in h_1 and h_2, i.e., p_1 and p_2, by forcing the proposition-level translation system to generate the projection observed in the corresponding hypothesis. The reachability of p_1 (respectively, p_2) and the decoder confidence in translating p as p_1 are used as features to estimate p_1 (p_2) accuracy. Similarly, for each argument a in each proposition p we generate its automatic projection on h_1 and h_2, i.e., a_1 and a_2. We force the argument-level decoder to translate a into a_1 and a_2, and use the respective reachability and translation confidence as features accounting for their appropriateness.

The best translation hypothesis (h_1 or h_2) is then selected according to the following decision function:

$$h^* = \underset{i \in \{0,1\}}{\arg\max} \sum_k \mathrm{f}_k(h_i, h_{j \neq i}, s) \qquad (1)$$

where each feature function $\mathrm{f}_k(\cdot,\cdot,\cdot)$ defines a comparison measure between its first two arguments, and returns 1 if the first argument is greater (better) than the second, and 0 otherwise. In short, the decision function selects the hypothesis that wins the highest number of comparisons.

The feature functions that we defined account for the following factors, the last three being evaluated once for each proposition in s: (1) Number of successfully translated propositions; (2) Average translation confidence for projected propositions; (3) Number of times that a proposition in h_i has higher confidence than the corresponding proposition in $h_{i \neq j}$; (4) Number of successfully translated arguments; (5) Average translation confidence for projected arguments; (6) Number of times that an argument in h_i has higher confidence than the corresponding argument in $h_{i \neq j}$.

With reference to Figure 2, the two translation hypotheses have been scored 4 (very good) and 2 (bad) by human annotators. The score assigned by the proposition decoder to p_1 is higher than p_2, hence comparisons (2) and (3) are won by h_1. According to the arguments decoder, h_1 does a better job at representing A0 and A2; h_2 is better at rendering A1, and $pred$ is a tie. Therefore, h_1 also prevails according to (6). Given the very high confidence assigned to the translation of A2 in h_1, the hypothesis also prevails in (5). In this case, (1) and (4) do not contribute to the decision as the two projections have the same coverage.

4 Evaluation

In this section, we present the results obtained by applying the proposed method to the task of ranking consistency, or pairwise ranking of alternative translations: that is, given a source sentence s, and two candidate translations h_1 and h_2, decide which one is a better translation for s. Pairwise ranking is a simplified setting for CE that is general enough to model the selection of the best translation among a finite set of alternatives. Even though it cannot measure translation quality in isolation, a reliable pairwise ranking model would be sufficient to solve many common practical CE problems, such as system ranking, user feedback filtering or hypotheses re-ranking.

4.1 Datasets

We ran our experiments on the human assessments released as part of the ACL Workshops on Machine Translations in 2007 (Callison-Burch et al., 2007), 2008 (Callison-Burch et al., 2008), 2009 (Callison-Burch et al., 2009) and 2010 (Callison-Burch et al., 2010). These datasets will be referred to as *wmtYY(t)* in the remainder, *YY* being the last two digits of the year of the workshop and $t = n$ for newswire data or $t = e$ for Europarl data. So, for example, *wmt08e* is the Europarl test set of the 2008 edition of the workshop. As our system is trained on Europarl data, newswire test sets are to be considered out-of-domain. All the experiments are relative to English to Spanish translations.

The *wmt08*, *wmt09* and *wmt10* datasets provide a ranking among systems within the range [1,5] (1 being the worst system, and 5 the best). The different datasets contain assessments for a different number of systems, namely: 11 for *wmt08(e)*, 10 for *wmt08(n)*, 9 for *wmt09* and 16 for *wmt10n*. Generally, multiple annotations are available for each annotated sentence. In all cases in which multiple assessments are available, we used the average of the assessments.

The *wmt07* dataset would be the most interesting of all, in that it provides separate assessments for the two main dimensions of translation quality, adequacy and fluency, as well as system rankings. Unluckily, the number of annotations in this dataset is very small, and after eliminating the ties the numbers are even smaller. As results on such small numbers would not be very representative, we decided not to include them in our evaluation.

We also evaluated on the dataset described in (Specia et al., 2010), which we will refer to as *specia*. As the system is based on Europarl data, it is to be considered an in-domain benchmark. The dataset includes results produced by four different systems, each translation being annotated by only one judge. Given the size of the corpus (the output of each system has been annotated on the same set of 4,000 sentences), this dataset is the most representative among those that we considered. It is also especially interesting for two other reasons: 1) systems are assigned a score ranging from 1 (*bad*) to 4 (*good as it is*) based on the number of edits required to produce a publication-ready translation. Therefore, here we have an absolute measure of translation accuracy, as opposed to relative rankings; 2) each system involved in the evaluation has very peculiar characteristics, hence they are very likely to generate quite different translations for the same input sentences.

4.2 Setup

Our model consists of four main components: an automatic semantic role labeler (to annotate source sentences); a lexical translation model (to generate the alignments required to map the annotations onto a translation hypothesis); a translation model for predicate-argument structures, to assign a score to projected annotations; and a translation model for role fillers, to assign a score to the projection of each argument.

To automatically label our training data with semantic roles we used the Swirl system[2] (Surdeanu and Turmo, 2005) with the bundled English models for syntactic and semantic parsing. On the CoNLL-2005 benchmark (Carreras and Màrquez, 2005), Swirl sports an F1-measure of 76.46. This figure drops to 75 for mixed data, and to 65.42 on out-of-domain data, which we can regard as a conservative estimate of the accuracy of the labeler on *wmt* benchmarks.

For all the translation tasks we employed the Moses phrase-based decoder[3] in a single-factor configuration. The `-constraint` command line parameter is used to force Moses to output the desired translation. For the English to Spanish lexical translation model, we used an already available model learned using all available *wmt10e* data.

To build the *proposition level* translation system, we first annotated all the English sentences from the *wmt10e* (en→es) training set with Swirl; then, we forced the lexical translation model to generate the alignments for the reference translations and projected the annotations on the target side. The process resulted in 2,493,476 parallel annotations. 5,000 annotations were held-out for model tuning. The training data was used to estimate a 5-gram language model and the translation model, which we later optimized on held-out data.

As for the *argument level* translator, we trained it on parallel word sequences spanning the same role in an annotation and its projection. Each such pair constitutes a training example for the argument translator, each argument representation being modeled independently from the others. With the same setup used for the proposition translator, we collected 4,578,480 parallel argument fillers from *wmt10e* en→es training data, holding out 20,000 pairs for model tuning.

[2] `http://www.surdeanu.name/mihai/swirl/`

[3] `http://www.statmt.org/moses/`

4.3 A note on recall

The main limitation of the model in its current implementation is its low recall. The translation model that we use to generate the alignments is mostly responsible for it. In fact, in approximately 35% of the cases the constrained translation model is not able to generate the required hypothesis. An obvious improvement would consist in using just an alignment model for this task, instead of resorting to translation, for instance following the approach adopted in (Esplà et al., 2011). It should also be noted that, while this component adds the interesting property of decoupling the measure from the system that produced the hypothesis, it is not strictly necessary in all those cases in which translation alignments are already available, e.g., for N-best re-ranking.

The second component that suffers from recall problems is the semantic role labeler, which fails in annotating sentences in approximately 6% of the remaining cases. These failures are by and large due to the lack of proper verbal predicates in the target sentence, and as such expose a limiting factor of the underlying model. In another 3% of the cases, an annotation is produced but it cannot be projected on the hypothesis, since the predicate word on the target side gets deleted during translation.

Another important consideration is that no measure for CE is conceived to be used in isolation, and our measure is no exception. In combination with others, the measure should only trigger when appropriate, i.e., when it is able to capture interesting patterns that are significant to discriminate translation quality. If it abstains, the other measures would compensate for the missing values. In this respect, we should also consider that not being able to produce a translation may be inherently considered an indicator of translation quality.

specia	Corr	Wrong	Und(%)	Acc(%)
$d = 1$	1076	656	14.26	62.12
$d = 2$	272	84	11.00	76.40
$d = 3$	30	8	13.64	78.95
$d \geq 1$	1378	748	13.72	**64.82**
$d \geq 2$	302	92	11.26	76.65
$d \geq 3$	30	8	13.64	78.95
wmt10n	Corr	Wrong	Und(%)	Acc(%)
$d = 1$	428	374	15.04	53.37
$d = 2$	232	196	18.01	54.21
$d = 3$	98	74	16.50	56.98
$d \geq 1$	784	664	16.20	**54.14**
$d \geq 2$	356	290	17.60	55.11
$d \geq 3$	124	94	16.79	56.88
wmt09n	Corr	Wrong	Und(%)	Acc(%)
$d = 1$	70	60	19.75	53.85
$d = 2$	30	40	20.45	42.86
$d = 3$	26	10	18.18	72.22
$d \geq 1$	134	116	19.87	**53.60**
$d \geq 2$	64	56	20.00	53.33
$d \geq 3$	34	16	19.35	68.00
wmt08n	Corr	Wrong	Und(%)	Acc(%)
$d = 1$	64	36	12.28	64.00
$d = 2$	26	24	19.35	52.00
$d = 3$	12	6	18.18	66.67
$d \geq 1$	104	70	14.71	**59.77**
$d \geq 2$	40	34	17.78	54.05
$d \geq 3$	14	10	14.29	58.33
wmt08e	Corr	Wrong	Und(%)	Acc(%)
$d = 1$	62	34	21.31	64.58
$d = 2$	40	30	10.26	57.14
$d = 3$	22	8	11.76	73.33
$d \geq 1$	134	80	15.75	**62.62**
$d \geq 2$	72	46	10.61	61.02
$d \geq 3$	32	16	11.11	66.67

Table 1: Results on five confidence estimation benchmarks. An *n* next to the task name (e.g. wmt08n) stands for a news (i.e. out of domain) corpus, whereas an *e* (e.g. wmt08e) stands for a Europarl (i.e. in domain) corpus. The *specia* corpus is in-domain.

4.4 Results

Table 1 lists, in each block of rows, pairwise classification accuracy results obtained on a specific benchmark. The benchmarks are sorted in order of reverse relevance, the largest benchmark (*specia*) being listed first. In each row, we show results obtained for different configurations in which the variable is the distance d between two assessment scores. So, for example, the row $d = 1$ accounts for all the comparisons in which the distance between scores is exactly one, while row $d \geq 2$ considers all the cases in which the distance is at least 2. For each test, the columns show: the number of correct (*Corr*) and wrong (*Wrong*) decisions, the percentage of undecidable cases (*Und*), i.e., the cases in which the scoring function cannot decide between the two hypotheses, and the accuracy of classification (*Acc*) measured without considering the unbreakable ties.

The accuracy for $d \geq 1$, i.e., on all the available annotations, is shown in bold.

First, we can observe that the results are above the baseline (an accuracy of 50% for evenly distributed binary classification) on all the benchmarks and for all configurations. The only outlier is *wmt09n* for $d = 2$, with an accuracy of 42.86%. Across the different datasets, results vary from promising (*specia* and *wmt08e*, where accuracy is generally above 60%) to mildly good (*wmt10n*), but across all the board the method seems to be able to provide useful clues for confidence estimation.

As expected, the accuracy of classification tends to increase as the difference between hypotheses becomes more manifest. In four cases out of six, the accuracy for $d = 3$ is above 60%, with the notable peaks on *specia*, *wmt09n* and *wmt08e* where it goes over 70% (on the first, it arrives almost at 80%). Unluckily, very few translations have very different quality (a measure of the difficulty of the task). Nevertheless, the general trend seems to support the reliability of the approach.

When we consider the results on the whole datasets (i.e., $d \geq 1$), pairwise classification accuracy ranges from 54% (for *wmt09n* and *wmt10n*, both out-of-domain), to 63-64% (for *specia* and *wmt08e*, both in-domain). Interestingly, the performance on *wmt08n*, which is also out-of-domain, is closer to in-domain benchmarks, i.e., 60%. These figures suggest that the method is consistently reliable on in-domain data, but also out-of-domain evaluation can benefit from its application. The difference in performance between *wmt08n* and the other out-of-domain benchmarks will be reason of further investigation as future work, as well as the drop in performance for $d = 2$ on three of the benchmarks.

5 Conclusions

We have presented a model to exploit the rich information encoded by predicate-argument structures for confidence estimation in machine translation. The model is based on a battery of translation systems, which we use to study the movement and the internal representation of propositions and arguments projected from source to target via automatic alignments. Our preliminary results, obtained on five different benchmarks, suggest that the approach is well grounded and that semantic annotations have the potential to be successfully employed for this task.

The model can be improved in many ways, its major weakness being its low recall as discussed in Section 4.3. Another area in which there is margin for improvement is the representation of predicate argument structures. It is reasonable to assume that different representations could yield very different results. Introducing more clues about the semantic content of the whole predicate argument structure, e.g., by including argument head words in the representation of the proposition, or considering a more fine-grained representation at the proposition level, could make it possible to assess the quality of a translation reducing the need to back-off to individual arguments. As for the representation of arguments, a first and straightforward improvement would be to train a separate model for each argument class, or to move to a factored model that would allow us to model explicitly the insertion of words or the overlap of argument words due to the projection.

Another important research direction involves the combination of this measure with already assessed metric sets for CE, e.g., (Specia et al., 2010), to understand to what extent it can contribute to improve the overall performance. In this respect, we would also like to move from a heuristic scoring function to a statistical model.

Finally, we would like to test the generality of the approach by designing other features based on the same "annotate, project, measure" framework, as we strongly believe that it is an effective yet simple way to combine several linguistic features for machine translation evaluation. For example, we would like to apply a similar framework to model the movement of chunks or POS sequences.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This research has been partially funded by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914).

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. ACL.

Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation*. ACL, Prague, Czech Republic.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce, editors. 2008. *Proceedings of the Third Workshop on Statistical Machine Translation*. ACL, Columbus, Ohio.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder, editors. 2009. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. ACL, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. ACL, Uppsala, Sweden.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, June. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Miquel Esplà, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Associtation for Machine Translation*.

Jesús Giménez and Lluís Màrquez. 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24:209–240. 10.1007/s10590-011-9088-7.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Chi-kiu Lo and Dekai Wu. 2010a. Evaluating machine translation utility via semantic role labels. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Chi-kiu Lo and Dekai Wu. 2010b. Semantic vs. syntactic vs. n-gram structure for machine translation evaluation. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 52–60, Beijing, China.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.*, 31(1):71–106.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. 2002. Inductive confidence machines for regression. In *AMAI'02*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. ACL.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Machine Translation Summit XII*, Ottawa, Canada.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Mihai Surdeanu and Jordi Turmo. 2005. Semantic role labeling using complete syntactic analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 221–224, Ann Arbor, Michigan, June. Association for Computational Linguistics.

S. Wold, A. Ruhe, H Wold, and W.J. Dunn. 1984. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. 5:735–743.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 13–16, Stroudsburg, PA, USA. ACL.