Match without a Referee: Evaluating MT Adequacy without Reference Translations

Yashar Mehdad **Matteo Negri** **Marcello Federico**
Fondazione Bruno Kessler, FBK-irst
Trento , Italy
{mehdad|negri|federico}@fbk.eu

Abstract

We address two challenges for automatic machine translation evaluation: a) avoiding the use of reference translations, and b) focusing on adequacy estimation. From an economic perspective, getting rid of costly hand-crafted reference translations (a) permits to alleviate the main bottleneck in MT evaluation. From a system evaluation perspective, pushing semantics into MT (b) is a necessity in order to complement the shallow methods currently used overcoming their limitations. Casting the problem as a cross-lingual textual entailment application, we experiment with different benchmarks and evaluation settings. Our method shows high correlation with human judgements and good results on all datasets without relying on reference translations.

1 Introduction

While syntactically informed modelling for statistical MT is an active field of research that has recently gained major attention from the MT community, work on integrating semantic models of adequacy into MT is still at preliminary stages. This situation holds not only for system development (most current methods disregard semantic information, in favour of statistical models of words distribution), but also for system evaluation. To realize its full potential, however, MT is now in the need of semantic-aware techniques, capable of complementing frequency counts with meaning representations.

In order to integrate semantics more deeply into MT technology, in this paper we focus on the evaluation dimension. Restricting our investigation to some of the more pressing issues emerging from this area of research, we provide two main contributions.

1. An automatic evaluation method that avoids the use of reference translations. Most current metrics are based on comparisons between automatic translations and human references, and reward lexical similarity at the n-gram level (*e.g.* BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006)). Due to the variability of natural languages in terms of possible ways to express the same meaning, reliable lexical similarity metrics depend on the availability of multiple hand-crafted (costly) realizations of the same source sentence in the target language. Our approach aims to avoid this bottleneck by adapting cross-lingual semantic inference capabilities and judging a translation only given the source sentence.

2. A method for evaluating translation adequacy. Most current solutions do not consistently reward translation adequacy (semantic equivalence between source sentence and target translation). The scarce integration of semantic information in MT, specifically at the multilingual level, led to MT systems that are "illiterate" in terms of semantics and meaning. Moreover, current metrics are often difficult to interpret. In contrast, our method targets the adequacy dimension, producing easily interpretable results (*e.g.* judgements in a 4-point scale).

Our approach builds on recent advances in cross-lingual textual entailment (CLTE) recognition, which provides a natural framework to address MT adequacy evaluation. In particular, we approach the problem as an application of CLTE where bi-

directional entailment between source and target is considered as evidence of translation adequacy. Besides avoiding the use of references, the proposed solution differs from most previous methods which typically rely on surface-level features, often extracted from the source or the target sentence taken in isolation. Although some of these features might correlate well with adequacy, they capture semantic equivalence only indirectly, and at the level of a probabilistic prediction. Focusing on a combination of surface, syntactic and semantic features, extracted from *both* source and target (*e.g.* "source-target length ratio", "dependency relations in common"), our approach leads to informed adequacy judgements derived from the actual observation of a translation given the source sentence.

2 Background

Some recent works proposed metrics able to approximately assess meaning equivalence between candidate and reference translations. Among these, (Giménez and Màrquez, 2007) proposed a heterogeneous set comprising overlapping and matching metrics, compiled from a rich set of variants at five different linguistic levels: lexical, shallow-syntactic, syntactic, shallow-semantic and semantic. More similar to our approach, (Padó et al., 2009) proposed semantic adequacy metrics that exploit feature representations motivated by Textual Entailment (TE). Both metrics, however, highly depend on the availability of multiple reference translations.

Early attempts to avoid reference translations addressed *quality estimation* (QE) by means of large numbers of source, target, and system-dependent features to discriminate between "good" and "bad" translations (Blatz et al., 2004; Quirk, 2004). More recently (Specia et al., 2010b; Specia and Farzindar, 2010; Specia, 2011) conducted a series of experiments using features designed to estimate translation post-editing effort (in terms of volume and time) as an indicator of MT output quality. Good results in QE have been achieved by adding linguistic information such as shallow parsing, POS tags (Xiong et al., 2010), or dependency relations (Bach et al., 2011; Avramidis et al., 2011) as features. However, in general these approaches do not distinguish between fluency (*i.e.* syntactic correctness of the output translation) and adequacy, and mostly rely on fluency-oriented features (*e.g.* "number of punctuation marks"). As a result, a simple surface form variation is given the same importance of a content word variation that changes the meaning of the sentence. To the best of our knowledge, only (Specia et al., 2011) proposed an approach to frame MT evaluation as an adequacy estimation problem. However, their method still includes many features which are not focused on adequacy, and often look either at the source or at the target in isolation (see for instance "source complexity" and "target fluency" features). Moreover, the actual contribution of the adequacy features used is not always evident and, for some testing conditions, marginal.

Our approach to adequacy evaluation builds on and extends the above mentioned works. Similarly to (Padó et al., 2009) we rely on the notion of textual entailment, but we cast it as a cross-lingual problem in order to bypass the need of reference translations. Similarly to (Blatz et al., 2004; Quirk, 2004), we try to discriminate between "good" and "bad" translations, but we focus on adequacy. To this aim, like (Xiong et al., 2010; Bach et al., 2011; Avramidis et al., 2011; Specia et al., 2010b; Specia et al., 2011) we rely on a large number of features, but focusing on source-target dependent ones, aiming at informed adequacy evaluation of a translation given the source instead of a more generic quality assessment based on surface features.

3 CLTE for adequacy evaluation

We address adequacy evaluation by adapting cross-lingual textual entailment recognition as a way to measure to what extent a source sentence and its automatic translation are semantically similar. CLTE has been proposed by (Mehdad et al., 2010) as an extension of textual entailment (Dagan and Glickman, 2004) that consists in deciding, given a text T and a hypothesis H *in different languages*, if the meaning of H can be inferred from the meaning of T.

The main motivation in approaching adequacy evaluation using CLTE is that an adequate translation and the source text should convey the same meaning. In terms of entailment, this means that an adequate MT output and the source sentence should entail each other (bi-directional entailment). Los-

ing or altering part of the meaning conveyed by the source sentence (*i.e.* having more, or different information in one of the two sides) will change the entailment direction and, consequently, the adequacy judgement. Framed in this way, CLTE-based adequacy evaluation methods can be designed to distinguish meaning-preserving variations from true divergence, regardless of reference translations.

Similarly to many monolingual TE approaches, CLTE solutions proposed so far adopt supervised learning methods, with features that measure to what extent the hypotheses can be mapped into the texts. The underlying assumption is that the probability of entailment is proportional to the number of words in H that can be mapped to words in T (Mehdad et al., 2011). Such mapping can be carried out at different word representation levels (*e.g.* tokens, lemmas, stems), possibly with the support of lexical knowledge in order to cross the language barrier between T and H (*e.g.* dictionaries, phrase tables).

Under the same assumption, since in the adequacy evaluation framework the entailment relation should hold in both directions, the mapping is performed both from the source to the target and vice-versa, building on features extracted from both sentences. Moreover, to improve over previous CLTE methods and boost MT adequacy evaluation performance, we explore the joint contribution of a number of lexical, syntactic and semantic features (Mehdad et al., 2012).

Concerning the features used, it's worth observing that the cost of implementing our approach (in terms of required resources and linguistic processors), and the need of reference translations are intrinsically different bottlenecks for MT. While the limited availability of processing tools for some language pairs is a "temporary" bottleneck, the acquisition of multiple references is a "permanent" one. The former cost is reducing over time due to the progress in NLP research; the latter represents a fixed cost that has to be eliminated. Similar considerations hold regarding the need of annotated data to develop our supervised learning approach. Concerning this, the cost of labelling source-target pairs with adequacy judgments is significantly lower compared to the creation of multiple references.

3.1 Features

In order to learn models for classification and regression we used the Support Vector Machine (SVM) algorithms implemented in the LIBSVM package (Chang and Lin, 2011) with a linear kernel and default parameters setting. Aiming at objective adequacy evaluation, our method limits the recourse to MT system-dependent features to reduce the bias of evaluating MT technology with its own core methods. The experiments described in the following sections are carried out on publicly available English-Spanish datasets, exploring the potential of a combination of surface, syntactic and semantic features. Language-dependent ones are extracted by exploiting processing tools for the two languages (part-of-speech taggers, dependency parsers and named entity recognizers), most of which are available for many languages.

Our feature set can be described as follows:

Surface Form (F) features consider the number of words, punctuation marks and non-word markers (*e.g.* quotations and brackets) in source and target, as well as their ratios (source/target and target/source), and the number of out of vocabulary terms encountered.

Shallow Syntactic (SSyn) features consider the number and ratios of common part-of-speech (POS) tags in source and target. Since the list of valid POS tags varies for different languages, we mapped English and Spanish tags into a common list using the FreeLing tagger (Carreras et al., 2004).

Syntactic (Syn) features consider the number and ratios of dependency roles common to source and target. To create a unique list of roles, we used the DepPattern (Otero and Lopez, 2011) package, which provides English and Spanish dependency parsers.

Phrase Table (PT) matching features are calculated as in (Mehdad et al., 2011), with a phrasal matching algorithm that takes advantage of a lexical phrase table extracted from a bilingual parallel corpus. The algorithm determines the number of phrases in the source (1 to 5-grams, at the level of

tokens, lemmas and stems) that can be mapped into target word sequences, and vice-versa. To build our English-Spanish phrase table, we used the Europarl, News Commentary and United Nations Spanish-English parallel corpora. After tokenization, the Giza++ (Och and Ney, 2000) and the Moses toolkit (Koehn et al., 2007) were respectively used to align the corpora and extract the phrase table. Although the phrase table was generated using MT technology, its use to compute our features is still compatible with a system-independent approach since the extraction is carried out without tuning the process towards any particular task. Moreover, our phrase matching algorithm integrates matches from overlapping n-grams of different size and nature (tokens, lemmas and stems) which current MT decoding algorithms cannot explore for complexity reasons.

Dependency Relation (DR) matching features target the increase of CLTE precision by adding syntactic constraints to the matching process. These features capture similarities between dependency relations, combining syntactic and lexical levels. We define a dependency relation as a triple that connects pairs of words through a grammatical relation. In a valid match, while the relation has to be the same, the connected words can be either the same, or semantically equivalent terms in the two languages. For example, *"nsubj (loves, John)"* can match *"nsubj (ama, John)"* and *"nsubj (quiere, John)"* but not *"dobj (quiere, John)"*. Term matching is carried out by means of a bilingual dictionary extracted from parallel corpora during PT creation. Given the dependency tree representations of source and target produced with DepPattern, for each grammatical relation *r* we calculate two DR matching scores as the number of matching occurrences of *r* in both source and target, respectively normalized by: *i)* the number of occurrences of *r* in the source, and *ii)* the number of occurrences of *r* in the target.

Semantic Phrase Table (SPT) matching features represent a novel way to leverage the integration of semantics and MT-derived techniques. Semantically enhanced phrase tables are used as a recall-oriented complement to the lexical PT matching features. SPTs are extracted from the same parallel corpora used to build lexical PTs, augmented with shallow semantic labels. To this aim, we first annotate the corpora with the FreeLing named-entity tagger, replacing named entities with general semantic labels chosen from a coarse-grained taxonomy (person, location, organization, date and numeric expression). Then, we combine the sequences of unique labels into one single token of the same label. Finally, we extract the semantic phrase table from the augmented corpora in the same way mentioned above. The resulting SPTs are used to map phrases between NE-annotated source-target pairs, similar to PT matching. SPTs offer three main advantages: *i)* semantic tags allow to match tokens that do not occur in the original parallel corpora used to extract the phrase table, *ii)* SPT entries are often short generalizations of longer original phrases, so the matching process can benefit from the increased probability of mapping higher order n-grams (*i.e.* those providing more contextual information), and *iii)* their smaller size has positive impact on system's efficiency, due to the considerable search space reduction.

4 Experiments and results

4.1 Datasets

Datasets with manual evaluation of MT output have been made available through a number of shared evaluation tasks. However, most of these datasets are not specifically annotated for adequacy measurement purposes, and the available adequacy judgements are limited to few hundred sentences for some language pairs. Moreover, most datasets are created by comparing reference translations with MT systems' output, disregarding the input sentences. Such judgements are hence biased towards the reference. Furthermore, the inter-annotator agreement is often low (Callison-Burch et al., 2007). In light of these limitations, most of the available datasets are *per se* not fully suitable for adequacy evaluation methods based on supervised learning, nor to provide stable and meaningful results. To partially cope with these problems, our experiments have been carried out over two different datasets:

- **16K:** *16.000* English-Spanish pairs, with Spanish translations produced by multiple MT

systems, annotated by professional translators with *quality* scores in a 4-point scale (Specia et al., 2010a).

- **WMT07:** *703* English-Spanish pairs derived from MT systems' output, with explicit *adequacy* judgements on a 5-point scale.

The two datasets present complementary advantages and disadvantages. On the one hand, although it is not annotated to explicitly capture meaning-related aspects of MT output, the quality oriented dataset has the main advantage of being large enough for supervised approaches. Moreover, it should allow to check the effectiveness of our feature set in estimating adequacy as a latent aspect of the more general notion of MT output quality. On the other hand, the smaller dataset is less suitable for supervised learning, but represents an appropriate benchmark for MT adequacy evaluation.

4.2 Adequacy and quality prediction

To experiment with our CLTE-based evaluation method minimizing overfitting, we randomized each dataset 5 times (D1 to D5), and split them into 80% for training and 20% for testing. Using different feature sets, we then trained and tested various regression models over each of the five splits, and computed correlation coefficients between the CLTE model predictions and the human gold standard annotations ([1-4] for quality, and [1-5] for adequacy).

16K quality-based dataset

In Table 1 we compare the Pearson's correlation coefficient of our SVM regression models against the results reported in (Specia et al., 2010b), calculated with the same three common MT evaluation metrics with a single reference: BLEU, TER and Meteor. For the sake of comparison, we also report the average quality correlation (QE) obtained by (Specia et al., 2010b) over the same dataset.[1]

The results show that the integration of syntactic and semantic information allows our adequacy-oriented model to achieve a correlation with human quality judgements that is always significantly higher[2] than the correlation obtained by the MT evaluation metrics used for comparison. As expected a considerable improvement over surface features is achieved by the integration of syntactic information. A further increase, however, is brought by the complementary contribution of SPT (*recall-oriented*, due to the higher coverage of semantics-aware phrase tables with respect to lexical PTs), and DR matching features (*precision-oriented*, due to the syntactic constraints posed to matching text portions). Although they are meant to capture meaning-related aspects of MT output, our features allow to outperform the results obtained by the generic quality-oriented features used by (Specia et al., 2010b), which do not discriminate between adequacy and fluency.[3] When dependency relations and phrase tables (both lexical and semantics-aware) are used in combination, our scores also outperform the average QE score. Finally, looking at the different random splits of the same dataset (D1 to D5), our correlation scores remain substantially stable, proving the robustness of our approach not only for adequacy, but also for quality estimation.

WMT07 adequacy-based dataset

In Table 2 we compare our regression model, obtained in the same way previously described, against three commonly used MT evaluation metrics (Callison-Burch et al., 2007). In this case, the reported results do not show the same consistency over the 5 randomized datasets (D1 to D5). However, it is worth pointing out that: *i)* the small dataset is particularly challenging to train models with higher correlation with humans, *ii)* our aim is checking how far we get using only adequacy-oriented features rather than outperforming BLEU/TER/Meteor at any cost, and *iii)* our results are not far from those achieved by metrics that rely on reference translations. Compared with Meteor, the correlation is even higher proving the effectiveness of the proposed method.

[1] We only show the average results reported in (Specia et al., 2010b), since the distributions of the 16K dataset is different from our randomized distribution.

[2] $p < 0.05$, calculated using the approximate randomization test implemented in (Padó, 2006).

[3] As reported in (Specia et al., 2010b), more than 50% (39 out of 74) of the features used is translation-independent (only source-derived features).

Features	D1	D2	D3	D4	D5	AVG
F	0.2506	0.2578	0.2436	0.2527	0.2443	0.25
SSyn+Syn	0.4387	0.4114	0.3994	0.4114	0.3793	0.41
F+SSyn+Syn	0.4215	0.4398	0.4059	0.4464	0.4255	0.428
F+SSyn+Syn+DR	0.4668	0.4602	0.4386	0.4437	0.4454	**0.451**
F+SSyn+Syn+DR+PT	0.4724	0.4715	0.4852	0.5028	0.4653	**0.48**
F+SSyn+Syn+DR+PT+SPT	0.4967	0.4802	0.4688	0.4894	0.4887	**0.485**
BLEU						0.2268
TER						0.1938
METEOR						0.2713
QE (Specia et al., 2010b)						0.4792

Table 1: Pearson's correlation between SVM regression and human quality annotation over 16K dataset.

Features	D1	D2	D3	D4	D5	AVG
F	0.10	0.03	0.04	0.10	0.14	0.083
SSyn+Syn	0.299	0.351	0.1834	0.2962	0.2417	0.274
F+SSyn+Syn	0.2648	0.2870	0.4061	0.3601	0.1327	0.29
F+SSyn+Syn+DR	0.3196	0.4568	0.2860	0.5057	0.4066	**0.395**
F+SSyn+Syn+DR+PT	0.3254	0.4710	0.3921	0.4599	0.3501	**0.40**
F+SSyn+Syn+DR+PT+SPT	0.3487	0.4032	0.4803	0.4380	0.3929	**0.413**
BLEU						0.466
TER						0.437
METEOR						0.357

Table 2: Pearson's correlation between SVM regression and human adequacy annotation over WMT07.

4.3 Multi-class classification

To further explore the potential of our CLTE-based MT evaluation method, we trained an SVM multi-class classifier to predict the exact adequacy and quality scores assigned by human judges. The evaluation was carried out measuring the accuracy of our models with 10-fold cross validation to minimize overfitting. As a baseline, we calculated the performance of the Majority Class (MjC) classifier proposed in (Specia et al., 2011), which labels all examples with the most frequent class among all classes. The performance improvement over the result obtained by the MjC baseline (Δ) has been calculated to assess the contribution of different feature sets.

16K quality-based dataset

The accuracy results reported in Table 3a show that also in this testing condition, syntactic and semantic features improve over surface form ones. Besides that, we observe a steady improvement over the MjC baseline (from 5% to 12%). This demonstrates the effectiveness of our adequacy-based features to predict exact quality scores in a 4-point scale, although this is a more challenging and difficult task than regression and binary classification. Such improvement is even more interesting considering that (Specia et al., 2010b) reported discouraging results with multi-class classification to predict quality scores. Moreover, while they claimed that removing target-independent features (*i.e.* those only looking at the source text) significantly degrades their QE performance, we achieved good results without using any of these features.

WMT07 adequacy-based dataset

As we can observe in Table 3b, all variations of adequacy estimation models significantly outperform the MjC baseline, with improvements rang-

Features	10-fold acc.	Δ
F	42.16%	5.16
Syn+SSyn	46.61%	9.61
F+Syn+SSyn	47.10%	10.10
F+Syn+SSyn+DR	47.26%	10.26
F+Syn+SSyn+DR+PT	48.15%	11.15
F+Syn+SSyn+DR+PT+SPT	**48.74**%	11.74
MjC	37%	-

(a) 16K dataset.

Features	10-fold acc.	Δ
F	50.07%	14.07
Syn+SSyn	54.19%	18.19
F+Syn+SSyn	54.34%	18.34
F+Syn+SSyn+DR	56.47%	20.47
F+Syn+SSyn+DR+PT	56.61%	20.61
F+Syn+SSyn+DR+PT+SPT	**56.75**%	20.75
MjC	36%	-

(b) WMT07 dataset

Table 3: Multi-class classification accuracy of the quality/adequacy scores.

Features	10-fold acc.	Δ
F	65.85%	11.85
Syn+SSyn	69.59%	15.59
F+Syn+SSyn	70.89%	16.89
F+Syn+SSyn+DR	71.39%	17.39
F+Syn+SSyn+DR+PT	71.92%	17.92
F+Syn+SSyn+DR+PT+SPT	**72.21**%	18.21
MjC	54%	-

(a) 16k dataset.

Features	10-fold acc.	Δ
F	83.24%	12.84
Syn+SSyn	83.67%	13.27
F+Syn+SSyn	84.31%	13.91
F+Syn+SSyn+DR	84.86%	14.46
F+Syn+SSyn+DR+PT	84.96%	14.56
F+Syn+SSyn+DR+PT+SPT	**85.20**%	14.80
MjC	70.4%	-

(b) WMT07 dataset.

Table 4: Accuracy of the binary classification into "good" or "adequate", and "bad" or "inadequate".

ing from 14% to 20%. Interestingly, although the dataset is small and the number of classes is higher (5-point scale), the improvement and overall results are better than those obtained on the 16K dataset. Such result confirms our hypothesis that adequacy-based features extracted from both source and target perform better on a dataset explicitly annotated with adequacy judgements. In addition, the improvement over the MjC baseline (Δ) of our best model is much higher (20%) than the one reported in (Specia et al., 2011) on adequacy estimation (6%). We are aware that their results are calculated over a dataset for a different language pair (*i.e.* English-Arabic) which brings up more challenges. However, our smaller dataset (700 vs 2580 pairs) and the higher number of classes (5 vs 4) compensate to some extent the difficulty of dealing with English-Arabic pairs.

4.4 Recognizing "good" vs "bad" translations

Last but not least, we considered the traditional scenario for quality and confidence estimation, which is a binary classification of translations into "good" and "bad" or, from the meaning point of view, "adequate" and "inadequate". Adequacy-oriented binary classification has many potential applications in the translation industry, ranging from the design of confidence estimation methods that reward meaning-preserving translations, to the optimization of the translation workflow. For instance, an "adequate" translation can be just post-edited in terms of fluency by a target language native speaker, without having any knowledge of the source language. On the other hand, an "inadequate" translation should be sent to a human translator or to another MT system, in order to reach acceptable adequacy. Effective automatic binary classification has an evident positive impact on such workflow.

16K quality-based dataset

We grouped the quality scores in the 4-point scale into two classes, where scores {1,2} are considered as "bad" or "inadequate", while {3,4} are taken as "good" or "adequate". We carried out learning and

classification using different sets of features with 10-fold cross validation. We also compared our accuracy with the MjC baseline, and calculated the improvement of each model (Δ) against it.

The results reported in Table 4a demonstrate that the accuracy of our models is always significantly superior to the MjC baseline. Moreover, also in this case there is a steady improvement using syntactic and semantic features over the results obtained by surface form features. Additionally, it is worth mentioning that the best model improvement over the baseline (Δ) is much higher (about 18%) than the improvement reported in (Specia et al., 2010b) over the same dataset (about 8%), considering the average score obtained with their data distribution. This confirms the effectiveness of our CLTE approach also in classifying "good" and "bad" translations.

WMT07 adequacy-based dataset

We mapped the 5-point scale adequacy scores into two classes, with {1,2,3} judgements assigned to the "inadequate" class, and {4,5} judgements assigned to the "adequate" class. The main motivation for this distribution was to separate the examples in a way that adequate translations are substantially acceptable, while inadequate translations present evident meaning discrepancies with the source.

The results reported in Table 4b show that the accuracy of the binary classifiers to distinguish between "adequate" and "inadequate" classes was significantly superior (up to about 15%) to the MjC baseline. We also notice that surface form features have a significant contribution to deal with the adequacy-oriented dataset, while the gain obtained using syntactic and semantic features (2%) is lower than the improvement observed in the 16K dataset. This might be due to the more unbalanced distribution of the classes which: *i)* leads to a high baseline, and *ii)* together with the small size of the WMT07 dataset, makes supervised learning more challenging. Finally, the improvement of all models (Δ) over the MjC baseline is much higher than the gain reported in (Specia et al., 2011) over their adequacy-oriented dataset (around 2%).

5 Conclusions

In the effort of integrating semantics into MT technology, we focused on automatic MT evaluation, investigating the potential of applying cross-lingual textual entailment techniques for adequacy assessment. The underlying assumption is that MT output adequacy can be determined by verifying that an entailment relation holds from the source to the target, and vice-versa. Within such framework, this paper makes two main contributions.

First, in contrast with most current metrics based on the comparison between automatic translations and multiple references, we avoid the bottleneck represented by the manual creation of such references.

Second, beyond current approaches biased towards fluency or general quality judgements, we tried to isolate the adequacy dimension of the problem, exploring the potential of adequacy-oriented features extracted from the observation of source and target.

To achieve our objectives, we successfully extended previous CLTE methods with a variety of linguistically motivated features. Altogether, such features led to reliable judgements that show high correlation with human evaluation. Coherent results on different datasets and classification schemes demonstrate the effectiveness of the approach and its potential for different applications.

Future works will address both the improvement of our adequacy evaluation method and its integration in SMT for optimization purposes. On one hand, we plan to explore new features capturing other semantic dimensions. A possible direction is to consider topic modelling techniques to measure the relatedness of source and target. Another interesting direction is to investigate the use of Wikipedia entity linking tools to support the mapping between source and target terms. On the other hand, we plan to explore the integration of our model as an error criterion in SMT system training.

Acknowledgments

This work has been partially supported by the CoSyne project (FP7-ICT-4-24853) and T4ME network of excellence (FP7-IST-249119), funded by the European Commission under the 7th Framework Programme. The authors would like to thank Hanna Bechara, Antonio Valerio Miceli Barone and Daniele Pighin for their contributions during the MT Marathon 2011.

References

E. Avramidis, M. Popovic, V. Vilar Torres, and A. Burchardt. 2011. Evaluate with Confidence Estimation: Machine Ranking of Translation Outputs using Grammatical Features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*.

N. Bach, F. Huang, and Y. Al-Onaizan. 2011. Goodness: A Method for Measuring Machine Translation Confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proceedings of the 20th international conference on Computational Linguistics (COLING '04)*. Association for Computational Linguistics.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT '07)*.

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

C.C. Chang and C.J. Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3).

I. Dagan and O. Glickman. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop of Learning Methods for Text Understanding and Mining*.

G. 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.

J. Giménez and L. Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07)*.

P. Koehn, 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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007)*.

Y. Mehdad, M. Negri, and M. Federico. 2010. Towards Cross-Lingual Textual Entailment. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*.

Y. Mehdad, M. Negri, and M. Federico. 2011. Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*.

Y. Mehdad, M. Negri, and M. Federico. 2012. Detecting Semantic Equivalence and Information Disparity in Cross-lingual Documents. In *Proceedings of the ACL'12*.

F.J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*.

P.G. Otero and I.G. Lopez. 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International journal of corpus linguistics*, 16(1).

S. Padó, M. Galley, D. Jurafsky, and C. D. Manning. 2009. Textual Entailment Features for Machine Translation Evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (StatMT '09)*.

S. Padó, 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation (ACL 2002. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

C.B. Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of LREC 2004*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*.

L. Specia and A. Farzindar. 2010. Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the AMTA-2010 Workshop, Bringing MT to the User: MT Research and the Translation Industry*.

L. Specia, N. Cancedda, and M. Dymetman. 2010a. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC10)*.

L. Specia, D. Raj, and M. Turchi. 2010b. Machine Translation Evaluation Versus Quality Estimation. *Machine translation*, 24(1).

L. Specia, N. Hajlaoui, C. Hallett, and W. Aziz. 2011. Predicting Machine Translation Adequacy. In *Proceedings of the 13th Machine Translation Summit (MT-Summit 2011)*.

L. Specia. 2011. Exploiting Objective Annotations for Minimising Translation Post-editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011)*.

D. Xiong, M. Zhang, and H. Li. 2010. Error Detection for Statistical Machine Translation Using Linguistic Features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Association for Computational Linguistics.