Exploiting Qualitative Information from Automatic Word Alignment for Cross-lingual NLP Tasks

José G.C. de Souza
FBK-irst,
University of Trento
Trento, Italy
desouza@fbk.eu

Miquel Esplà-Gomis
Universitat d'Alacant
Alacant, Spain
mespla@dlsi.ua.es

Marco Turchi
FBK-irst
Trento, Italy
turchi@fbk.eu

Matteo Negri
FBK-irst
Trento, Italy
negri@fbk.eu

Abstract

The use of automatic word alignment to capture sentence-level semantic relations is common to a number of cross-lingual NLP applications. Despite its proved usefulness, however, word alignment information is typically considered from a quantitative point of view (*e.g.* the number of alignments), disregarding qualitative aspects (the importance of aligned terms). In this paper we demonstrate that integrating qualitative information can bring significant performance improvements with negligible impact on system complexity. Focusing on the cross-lingual textual entailment task, we contribute with a novel method that: *i)* significantly outperforms the state of the art, and *ii)* is portable, with limited loss in performance, to language pairs where training data are not available.

1 Introduction

Meaning representation, comparison and projection across sentences are major challenges for a variety of cross-lingual applications. So far, despite the relevance of the problem, research on multilingual applications has either circumvented the issue, or proposed partial solutions.

When possible, the typical approach builds on the reduction to a monolingual task, burdening the process with dependencies from machine translation (MT) components. For instance, in cross-lingual question answering and cross-lingual textual entailment (CLTE), intermediate MT steps are respectively performed to ease answer retrieval/presentation (Parton, 2012; Tanev et al., 2006) and semantic inference (Mehdad et al., 2010). Direct solutions that avoid such pivoting strategies typically exploit similarity measures that rely on bag-of-words representations. As an example, most supervised approaches to MT quality estimation (Blatz et al., 2003; Callison-Burch et al., 2012) and CLTE (Wäschle and Fendrich, 2012) include features that consider the amount of equivalent terms that are found in the input sentence pairs. Such simplification, however, disregards the fact that semantic equivalence is not only proportional to the number of equivalent terms, but also to their importance. In other words, instead of checking *what* of a given sentence can be found in the other, current approaches limit the analysis to *the amount* of lexical elements they share, under the rough assumption that the more the better.

In this paper we argue that:
(1) Considering qualitative aspects of word alignments to identify sentence-level semantic relations can bring significant performance improvements in cross-lingual NLP tasks.
(2) Shallow linguistic processing techniques (often a constraint in real cross-lingual scenarios due to limited resources availability) can be leveraged to set up portable solutions that still outperform current bag-of-words methods.

To support our claims we experiment with the CLTE task, which allows us to perform exhaustive comparative experiments due to the availability of comparable benchmarks for different language pairs. In the remainder of the paper, we:
(1) Prove the effectiveness of our method over datasets for four language combinations;
(2) Assess the portability of our models across languages in different testing conditions.

2 Objectives and Method

We propose a supervised learning approach for identifying and classifying semantic relations between two sentences T_1 and T_2 written in different languages. Beyond semantic equivalence, which is relevant to applications such as MT quality es-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 771–776,
Sofia, Bulgaria, August 4-9 2013.

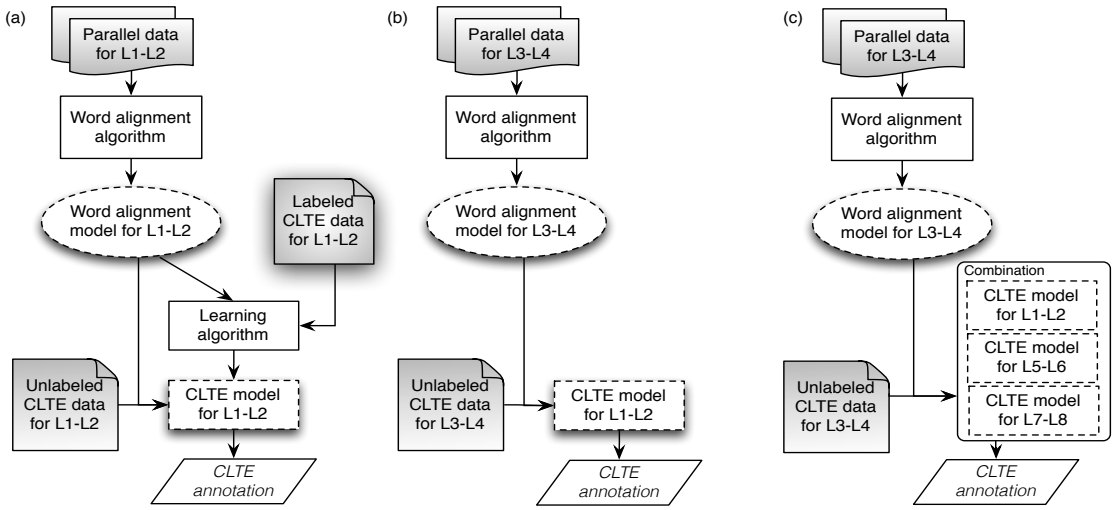

Figure 1: System architecture in different training/evaluation conditions. (a): parallel data and CLTE labeled data are available for language pair L1-L2. (b): the L1-L2 CLTE model is used to cope with the unavailability of labeled data for L3-L4. (c): the same problem is tackled by combining multiple models.

timation (Mehdad et al., 2012b),[1] we aim to capture a richer set of relations potentially relevant to other tasks. For instance, recognizing unrelatedness, forward and backward entailment relations, represents a core problem in cross-lingual document summarization (Lenci et al., 2002) and content synchronization (Monz et al., 2011; Mehdad et al., 2012a). CLTE, as proposed within the SemEval evaluation exercises (Negri et al., 2012; Negri et al., 2013), represents an ideal framework to evaluate such capabilities. Within this framework, our goal is to automatically identify the following entailment relations between T_1 and T_2: *forward* ($T_1 \rightarrow T_2$), *backward* ($T_1 \leftarrow T_2$), *bidirectional* ($T_1 \leftrightarrow T_2$) and *no_entailment*.

Our approach (see Figure 1) involves two core components: *i)* a word alignment model, and *ii)* a CLTE classifier. The former is trained on a parallel corpus, and associates equivalent terms in T_1 and T_2. The information about word alignments is used to extract quantitative (amount and distribution of the alignments) and qualitative features (importance of the aligned terms) to train the CLTE classifier. Although in principle both components need training data (respectively a parallel corpus and labeled CLTE data), our goal is to develop a method that is also portable across languages. To this aim, while the parallel corpus is necessary to train the word aligner for any language pair we want to deal with, the CLTE classifier can be designed to learn from features that capture language independent knowledge.[2] This allows us to experiment in different testing conditions, namely: *i)* when CLTE training data are available for a given language pair (Figure 1*a*), and *ii)* when CLTE training data are missing, and a model trained on other language pairs has to be reused (Figure 1*b-c*).

Features. Considering word alignment information, we extract three different groups of features: **AL**, **POS**, and **IDF**.

The **AL** group provides *quantitative* information about the aligned/unaligned words in each sentence T_* of the pair. These features are:

1. proportion of aligned words in T_*. We use this indicator as our baseline (**B** henceforth);
2. number of sequences of unaligned words, normalized by the length of T_*;
3. length of the longest *a*) sequence of aligned words, and *b*) sequence of unaligned words, both normalized by the length of T_*;
4. average length of *a*) the aligned word sequences, and *b*) the unaligned word sequences;
5. position of *a*) the first unaligned word, and *b*) the last unaligned word, both normalized by the lenght of T_*;
6. proportion of word n-grams in T_* containing only aligned words (the feature was com-

[1] A translation has to be semantically equivalent to the source sentence.

[2] For instance, the fact that aligning all nouns and the most relevant terms in T_1 and T_2 is a good indicator of semantic equivalence.

puted separately for values of $n = 1 \ldots 5$).

The **POS** group considers the part of speech (PoS) of the words in T_* as a source of *qualitative* information about their importance. To compute these features we use the TreeTagger (Schmid, 1995), manually mapping the fine-grained set of assigned PoS labels into a more general set of tags (P) based on the *universal PoS tag set* by Petrov et al. (2012). POS features differentiate between **aligned words** (words in T_1 that are aligned to one or more words in T_2) and **alignments** (the edges connecting words in T_1 and T_2). Features considering the aligned words in T_* are:

7. for each PoS tag $p \in P$, proportion of aligned words in T_* tagged with p;
8. proportion of words in T_1 aligned with words with the same PoS tag in T_2 (and vice-versa);
9. for each PoS tag $p \in P$, proportion of words in T_1 tagged as p which are aligned to words with the same tag in T_2 (and vice-versa).

Features considering the alignments are:

10. proportion of alignments connecting words with the same PoS tag p;
11. for each PoS tag $p \in P$, proportion of alignments connecting two words tagged as p.

IDF, the last feature, uses the inverse document frequency (Salton and Buckley, 1988) as another source of *qualitative* information under the assumption that rare words (and, therefore, with higher IDF) are more informative:

12. summation of all the IDF scores of the aligned words in T_* over the summation of the IDF scores of all words in T_*.

3 Experiments

Our experiments cover two different scenarios. First, the typical one, in which the CLTE model is trained on labeled data for the same pair of languages L_1–L_2 of the test set. Then, simulating the less favorable situation in which labeled training data for L_1–L_2 are missing, we investigate the possibility to use existing CLTE models trained on labeled data for a different language pair L_3–L_4.

The SemEval 2012 CLTE datasets used in our experiments are available for four language pairs: Es–En, De–En, Fr–En, and It–En. Each dataset was created with the crowdsourcing-based method described in Negri et al. (2011), and consists of 1000 T_1–T_2 pairs (500 for training, 500 for test).

To train the word alignment models we used the Europarl parallel corpus (Koehn, 2005), concatenated with the News Commentary corpus[3] for three language pairs: De–En (2,079,049 sentences), Es–En (2,123,036 sentences), Fr–En (2,144,820 sentences). For It–En we only used the parallel data available in Europarl (1,909,115 sentences) since this language pair is not covered by the News Commentary corpus. IDF values for the words in each language were calculated on the monolingual part of these corpora, using the average IDF value of each language for unseen terms.

To build the word alignment models we used the MGIZA++ package (Gao and Vogel, 2008). Experiments have been carried out with the *hidden Markov model* (HMM) (Vogel et al., 1996) and *IBM models* 3 and 4 (Brown et al., 1993).[4] We also explored three symmetrization techniques (Koehn et al., 2005): *union*, *intersection*, and *grow-diag-final-and*. A greedy feature selection process on training data, with different combinations of word alignment models and symmetrization methods, indicated *HMM*/*intersection* as the best performing combination. For this reason, all our experiments use this setting.

The SVM implementation of Weka (Hall et al., 2009) was used to build the CLTE model.[5] Two binary classifiers were trained to separately check $T_1 \rightarrow T_2$ and $T_1 \leftarrow T_2$, merging their output to obtain the 4-class judgments (*e.g.* `yes`/`yes`=bidirectional, `yes`/`no`=forward).

3.1 Evaluation with CLTE training data

Figure 2 shows the accuracy obtained by the different feature groups.[6] For the sake of comparison, state-of-the-art results achieved for each language combination at SemEval 2012 are also reported. As regards Es–En (63.2% accuracy) and De–En (55.8%), the top scores were obtained by the system described in (Wäschle and Fendrich, 2012), where a combination of binary classifiers for each entailment direction is trained with a mix-

[3] `http://www.statmt.org/wmt11/translation-task.html#download`

[4] Five iterations of HMM, and three iterations of IBM models 3 and 4 have been performed on the training corpora.

[5] The polynomial kernel was used with parameters empirically estimated on the training set ($C = 2.0$, and $d = 1$)

[6] In Figures 2 and 3, the "*" indicates statistically significant improvements over the state of the art at $p \leq 0.05$, calculated with approximate randomization (Padó, 2006).

ture of monolingual (*i.e.* with the input sentences translated in the same language using Google Translate[7]) and cross-lingual features. Although such system exploits word-alignment information to some extent, this is only done at quantitative level (*e.g.* number of unaligned words, percentage of aligned words, length of the longest unaligned subsequence). As regards It–En, the state of the art (56.6%) is represented by the system described in (Jimenez et al., 2012), which uses a pure pivoting method (using Google Translate) and adaptive similarity functions based on "soft" cardinality for flexible term comparisons. The two systems obtained the same result on Fr–En (57.0%).

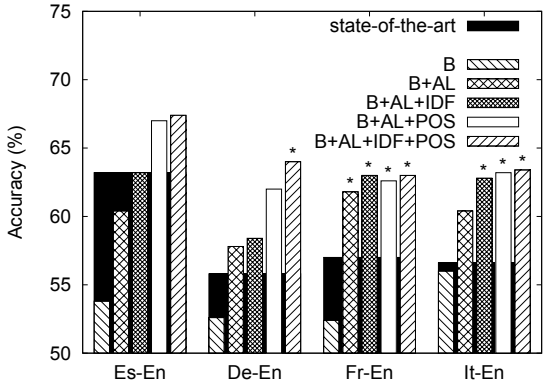

Figure 2: Accuracy obtained by each feature group on four language combinations.

As can be seen in Figure 2, the combination of *all* our features outperforms the state of the art for each language pair. The accuracy improvement ranges from 6.6% for Es–En (from 63.2% to 67.4%) to 14.6% for De–En (from 55.8% to 64%). Except for Es–En, that has very competitive state-of-the-art results, the combination of AL with POS or IDF feature groups always outperforms the best systems. Furthermore, the performance increase with qualitative features (POS and IDF) shows coherent trends across all language pairs. It is worth noting that, while we rely on a pure cross-lingual approach, both the state-of-the-art CLTE systems include features from the translation of T_1 into the language of T_2. For De–En, quantitative features alone achieve lower results compared to the other languages. This can be motivated by the higher difficulty in aligning De–En pairs (this hypothesis is supported by the fact that the average number of alignments per sentence pair is 18 for De–En, and >22 for the other combinations). Nevertheless, qualitative features lead to results comparable with the other language pairs.

The selection of the best performing features for each language pair produces further improvements of varying degrees in Es–En (from 67.4% to 68%), De–En (64% – 64.8%) and It–En (63.4% – 66.8%), while performance remains stable for Fr–En (63%). All these configurations include the IDF feature (12) and the proportion of aligned words for each PoS category (7), proving the effectiveness of qualitative word alignment features.

The fact that HMM/intersection is the best combination of alignment model and symmetrization method is interesting, since it contradicts the general notion that IBM models 3 and 4 perform better than HMM (Och and Ney, 2003). A possible explanation is that, while word alignment models are usually trained on parallel corpora, the majority of CLTE sentence pairs are not parallel. In this setting, where producing reliable alignments is more difficult, IBM models are less effective for at least two reasons. First, including a word fertility model, IBM 3 and 4 limit (typically to the half of the source sentence length) the number of target words that can be aligned with the `null` word. Therefore, when such limit is reached, these models tend to force low probability, hence less reliable, word alignments. Second, in IBM model 4, the larger distortion limit makes it possible to align distant words. In the case of non-parallel sentences, this often results in wrong or noisy alignments that affect final results. For these reasons, CLTE data seem more suitable for the simpler and more conservative HMM model, and a precision-oriented symmetrization method like intersection.

3.2 Evaluation without CLTE training data

The goal of our second round of experiments is to investigate if, and to what extent, our approach can be considered as language-independent. Confirming this would allow to reuse models trained for a given language pair in situations where CLTE training data is missing. This is a rather realistic situation since, while bitexts to train word aligners are easier to find, the availability of labeled CLTE data is far from being guaranteed.

Our experiments have been carried out, over the same SemEval datasets, with two methods that do not use labeled data for the target language combination. The first one (method *b* in Figure 1) uses a CLTE model trained for a language pair L_1–L_2 for which labeled training data are avail-

[7] `http://translate.google.com/`

able, and applies this model to a language pair L_3–L_4 for which only parallel corpora are available. The second method (*c* in Figure 1) addresses the same problem, but exploits a combination of CLTE models trained for different language pairs. For each test set, the models trained for the other three language pairs are used in a voting scheme, in order to check whether they can complement each other to increase final results.

All the experiments have been performed using the best CLTE model for each language pair, comparing results with those presented in Section 3.1.

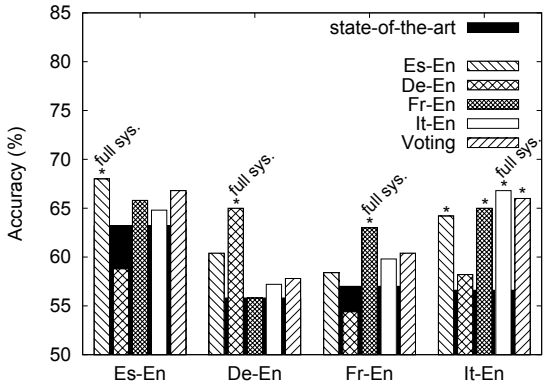

Figure 3: Accuracy obtained by reusing CLTE models (alone and in a voting scheme).

As shown in Figure 3, reusing models for a new language pair leads to results that still outperform the state of the art.[6] Remarkably, when used for other language combinations, the Es–En, It–En, and Fr–En models always lead to results above, or equal to the state of the art. For similar languages such as Spanish, French, and Italian, the accuracy increase over the state of the art is up to 14.8% (from 56.6% to 65.0%) and 13.4% (from 56.6% to 64.2%) when the Fr–En and Es–En models are respectively used to label the It–En dataset. Although not always statistically significant and below the performance obtained in the ideal scenario where CLTE training data are available (*full sys.*), such improvements suggest that our features can be re-used, at least to some extent, across different language settings. As expected, the major incompatibilities arise between German and the other languages due to the linguistic differences between this language and the others. However, it is interesting to note that: *i)* at least in one case (*i.e.* when tested on It–En) the De–En model still achieves results above the state of the art, and *ii)* on the De–En evaluation setting the worst model (Fr–En) still achieves state of the art results.

The results obtained with the voting scheme suggest that our models can complement each other when used on a new language pair. Although statistically significant only over It–En data, voting results both outperform the state of the art and the results achieved by single models.

4 Conclusion

We investigated the usefulness of qualitative information from automatic word alignment to identify semantic relations between sentences in different languages. With coherent results in CLTE, we demonstrated that features considering the importance of aligned terms can successfully integrate the quantitative evidence (number and proportion of aligned terms) used by previous supervised learning approaches. A study on the portability across languages of the learned models demonstrated that word alignment information can be exploited to train reusable models for new language combinations where bitexts are available but CLTE labeled data are not.

Acknowledgments

This work has been partially supported by the EC-funded projects CoSyne (FP7-ICT-4-248531) and MateCat (ICT-2011.4.2–287688), and by Spanish Government through projects TIN2009-14009-C02-01 and TIN2012-32615.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation. Summer workshop final report, JHU/CLSP.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT'12)*, pages 10–51, Montréal, Canada.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, USA.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10–18.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality + ML: Learning Adaptive Similarity Functions for Cross-lingual Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 684–688, Montréal, Canada.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Philip Koehn. 2005. Europarl: a Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Alessandro Lenci, Roberto Bartolini, Nicoletta Calzolari, Ana Agua, Stephan Busemann, Emmanuel Cartier, Karine Chevreau, and José Coch. 2002. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI Project. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1464–1471, Las Palmas de Gran Canaria, Spain.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual Entailment. In *Proceedings of the Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 321–324, Los Angeles, California, USA.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012a. Detecting Semantic Equivalence and Information Disparity in Cross–lingual Documents. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 120–124, Jeju Island, Korea.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012b. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, Montréal, Canada.

Christoph Monz, Vivi Nastase, Matteo Negri, Angela Fahrni, Yashar Mehdad, and Michael Strube. 2011. CoSyne: a Framework for Multilingual Content Synchronization of Wikis. In *Proceedings of WikiSym 2011, the International Symposium on Wikis and Open Collaboration*, pages 217–218, Mountain View, California, USA.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, Scotland.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 Task 8: Cross-Lingual Textual Entailment for Content Synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 Task 8: Cross-Lingual Textual Entailment for Content Synchronization. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, GA.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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

Kristen Parton. 2012. *Lost and Found in Translation: Cross-Lingual Question Answering with Result Translation*. Ph.D. thesis, Columbia University.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey.

Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Hristo Tanev, Milen Kouylekov, Bernardo Magnini, Matteo Negri, and Kiril Simov. 2006. Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005. *Accessing Multilingual Information Repositories*, pages 390–399.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. In *Proceedings of the 16th International Conference on Computational Linguistics (ACL'96)*, pages 836–841, Copenhagen, Denmark.

Katharina Wäschle and Sascha Fendrich. 2012. HDU: Cross-lingual Textual Entailment with SMT Features. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 467–471, Montréal, Canada.