# Document Translation Retrieval Based on Statistical Machine Translation Techniques

Felipe Sánchez-Martínez, Rafael C. Carrasco
*Departament de Llenguatges i Sistemes Informàtics*
*Universitat d'Alacant, E-03071 Alacant, Spain*
{fsanchez,carrasco}@dlsi.ua.es

**Abstract**

We compare different strategies to apply statistical machine translation techniques in order to retrieve documents which are a plausible translation of a given source document. Finding the translated version of a document is a relevant task, for example, when building a corpus of parallel texts that can help to create and to evaluate new machine translation systems.

In contrast to the traditional settings in cross-language information retrieval tasks, in this case both the source and the target text are long and, thus, the procedure used to select what words or phrases will be included in the query has a key effect on the retrieval performance. In the statistical approach explored here, both the probability of the translation and the relevance of the terms are taken into account in order to build an effective query.

## 1 Introduction

Imagine the following scenario: a journal editor asks Susan to translate into Spanish a well-known text, such as the German version of the A. Einstein's research article "Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes" (Einstein, 1911). Before she starts working, she would like to know if there are already digital versions of this text in Spanish available in her digital library. For that purpose she uses a search engine to look for instances of this article. When she introduces the title in Spanish —which is usually translated as "Sobre la influencia de la gravedad en la propagación de la luz"—, she finds thousands of courses, dissertations and biographical excerpts which include references to the original paper but cannot find, even after browsing tens of results, the desired document. She then realizes that it would be more effective to translate some words or phrases from the German content into Spanish; however, this task requires




The definitive version was published in Applied Artificial Intelligence, Volume 25 Issue 5, May 2011. doi:10.1080/08839514.2011.559906 (http://dx.doi.org/10.1080/08839514.2011.559906)

some additional, non-automatic, work. For example, the title of the first section in the aforementioned article starts with the word "Hypothese" which is hardly effective to discriminate results when looking for a scientific document. The title of the first section also includes the more specific term "des Gravitationfeldes" but, it is probable that one finds it sometimes translated as "del campo gravitatorio" and sometimes as "del campo gravitacional". Clearly, some of the terms in the source text are too frequent to be useful to filter documents while some other terms allow for multiple translations and one cannot safely predict in what form they will appear in the target document.

Similar difficulties appear when a researcher wants to use the *world wide web* (WWW) to build a large translation corpus. The WWW can be regarded as a huge corpus containing millions of texts of variable quality (Kilgarriff and Grefenstette, 2003) and, thus, a collection of bitexts (a *bitext* is composed of versions in two different languages of a given text) can be built by finding pairs of documents in the web which are mutual translations. In order to identify automatically pairs of *Uniform Resource Locators* (URL) whose contents are bitext candidates, some approaches (Resnik and Smith, 2003) take into account the similitude of the URLs, the textual content of the pages and, to some extent, the structure of the text provided by the HTML tags (Sánchez-Villamil et al., 2006). Here, we will explore a complementary approach: after a collection of texts in the source language is compiled, documents which are a possible translation of the source texts are sought for. Again, the task is hindered by the variability of human translation, and translating (manually or automatically) a large fragment of the source document will usually return no result when a search engine is used to locate the document containing the translated content. Therefore, a reasonable approach would be to look for relevant words or expressions with a high probability of being contained in the translated document.

The *statistical machine translation* (SMT) framework provides methods to measure both the relevance of a phrase and the safety (that is, the relative frequency) of its translations (Brown et al., 1990, 1993; Koehn, 2010). The application of SMT techniques to *information retrieval* tasks has been explored before. For instance, the relevance of a document can be estimated through a probabilistic model if the query is considered to be a translation of the document (Berger and Lafferty, 1999; Federico and Bertoldi, 2002). Although this approach cannot aim at describing how users actually create queries, it generates promising empirical results. In *cross-language information retrieval*, SMT methods have been often used to address the inherent ambiguity of translation (Kishida, 2005), although using all possible translations for searching provides good average precision (Hiemstra and de Jong, 1999). In some cases, documents are translated before indexing them (Grefenstette, 1998), but query translation is often preferred because it requires fewer computational resources (McCarley, 1999). However, when looking for the translation of a document, both the source and the target texts are long and, thus, statistical disambiguation is not effective —as the probability that

the disambiguation is totally correct quickly vanishes— unless the query is shortened.

In the following we present a probabilistic description of the task of retrieving a document translation (section 2) and the corpora and tools used to evaluate this approach (section 3). In section 4, different methods to exploit the information provided by statistical machine translation tools are compared and conclusions are presented in section 5.

# 2 Model description

It will be assumed in the following that a large multilingual collection of documents $\Omega$ is provided and that it can be uniquely partitioned into a finite number of subsets $\Omega_L$, each one containing only those documents written in a given language $L$. The objective is to retrieve the documents in $\Omega_T$, where $T$ is the target language, which are likely to be the (human) translation of a given document written in a source language $S$. In the statistical machine translation framework, every source document $X$ consist of a number $M$ of segments $x_1 x_2 \cdots x_M$ that can be translated independently. Depending on the particular approach, a segment is always a single word or it can also be a multi-word expression —often called a *phrase* in the SMT literature (Zens et al., 2002; Koehn et al., 2003; Koehn, 2010) even if they are not necessarily a well-formed syntactic constituent. The result of translating a document $X = x_1 \cdots x_M \in \Omega_S$ is a new document $Y = y_1 \cdots y_N \in \Omega_T$ containing the translation of all the source segments (although some of them may be translated into an empty phrase), perhaps, after some reordering —that is, the translation is not necessarily monotonic. If $y_n$ in $Y$ is the result of translating $x_m$ in $X$, both segments are said to be *aligned*. For every pair of segments $(x_m, y_n)$ in the source and target languages, the SMT system provides:

- a subset $\tau(x_m)$ of possible translations for $x_m$;
- the probability $p_{ST}(y_n|x_m)$ that $x_m$ is translated as $y_n$;
- the reordering costs, given by the alignment probabilities $p_A(m, n, x_m, y_n)$ which depend, usually, on the distance $n - m$ and, optionally, on the pair components; and
- the probability $p_T(Y)$ that $Y = y_1 y_2 \ldots y_N$ is a document in $\Omega_T$, called the *target-language model*.

These probability functions are estimated using a collection of bitexts —the *training subset*— and, then, used to generate and evaluate possible translations of the source documents contained in a different collections of translations —the *test subset*. For example, a traditional formulation (Brown et al., 1990) breaks documents into sentences and uses the expectation-maximization algorithm to obtain the translation and

alignment probabilities (Brown et al., 1993) that maximize, for every pair of sentences $(f, e)$, the product between the probability of $e$ in the target language and the probability that $e$ is translated as $f$. However, recent SMT systems (Zens et al., 2002; Och and Ney, 2002; Koehn et al., 2003) build a log-linear combination of a number of *feature functions*

$$\exp \sum_k \lambda_k h_k(f_1 \cdots f_M, e_1 \cdots e_N) \tag{1}$$

where the parameters $\lambda_k$ are selected to optimize the translation results obtained over a disjoint collection of bitexts —the *development subset*. This optimization is traditionally based on BLEU —*bilingual evaluation understudy* (Papineni et al., 2002)—, an approximate indicator (Callison-Burch et al., 2006) which can be estimated automatically and does not require additional supervision by experts of the translation results.

The feature functions in Eq. (1) are logarithms of probabilities (Och and Ney, 2002), but other distances can also be included. Typical feature functions are source-to-target and target-to-source phrase translation probabilities ($p_{ST}$ and $p_{TS}$), source-to-target and target-to-source word translation probabilities (equivalent to a probabilistic bilingual dictionary), reordering costs, the output length (measured by the number of words), the number of phrases or segments used, and the target-language model ($p_T$).

For every document $X \in \Omega_S$, $\tau(X) \subset \Omega_T$ will denote the subset of documents in the target language $T$ which are a translation of $X$. Moreover, if $Y = y_1 \cdots y_N$, we will write $w \in Y$ if there exists $n \leq N$ such that $y_n = w$. Then, the probability that a randomly selected document $Y \in \Omega_T$ is a translation of $X$ when $Y$ contains a segment identical to $w$ can be written as

$$P(Y \in \tau(X) | w \in Y) = \frac{P(w \in Y | Y \in \tau(X))}{P(w \in Y)} \, P(Y \in \tau(X)) \tag{2}$$

where the last factor $P(Y \in \tau(X))$ does not depend on $w$ but only on the properties of the collections $\Omega_S$ and $\Omega_T$. The factor $P(w \in Y)$ in the previous equation can be interpreted as the *document frequency* (Manning et al., 2008, p. 118), that is, the number of documents in the collection that contain a term. Under the assumption that each segment in $Y$ is generated as the translation of a segment in $X$ one has

$$P(w \in Y | Y \in \tau(X)) = 1 - \prod_{m=1}^{M} (1 - P(w \in \tau(x_m)) \tag{3}$$

where $P(w \in \tau(x_m))$ is the probability that the translation of $x_m$ contains $w$, which is given by

$$P(w \in \tau(x_m)) = \sum_{y \in \tau(x_m) : w \in y} p_{ST}(y | x_m)$$

Of course, Eq. (3) is only approximate if segments can be multi-word expressions because the probability that the term is contained in the merge of two consecutive segments is neglected. Eq. (2) is consistent with the intuition that the document translation is more likely to be retrieved if the terms used in the query have simultaneously a high probability of being in the translated document and a low overall probability of being in the collection.

# 3 Resources and experimental settings

In order to test the approach proposed here, the following external resources have been used:

1. A parallel corpus, widely-used to evaluate and compare the results obtained with machine translation systems.

2. A standard, phrase-based statistical machine translation system together with a language modelling toolkit to find and evaluate the translations of every source document used for testing.

3. A text search engine to index and retrieve the target documents.

The *Europarl* corpus —*European Parliament Proceedings Parallel Corpus* (Koehn, 2005)— has been used to train the probabilistic models for several language pairs (English–Spanish, English–French and English–German) and also to evaluate their performance. This corpus consists of the textual transcriptions of the debates held at the European Parliament, organized by date. The content is split into chapters containing the discussion of a specific topic. As these chapters are a suitable target for the retrieval experiments, they play the role of documents in the experiments. The chapters from April 1998 to October 2006 have been selected as training set and the chapters from April 1996 to March 1998 as test set. Additional chapters[1] have been used as small development subset. Some files in the *Europarl* corpus are not split into chapters, leading to a few documents which are much larger than the average. In order to keep the size of the documents in the test subset homogeneous, the date ranges have been carefully selected so that these large files are all included in the training subset. Furthermore, the first and last chapter of each session have been removed from the test subset as they are usually short statements only announcing the resumption or the adjournment of a session. Table 1 summarizes the number of chapters, sentences and words in the training and test subsets.

[1]The 2,000 sentences per language pair released for the EACL 2009 workshop on SMT, which are available at `http://www.statmt.org/wmt09/additional-dev.tgz`.

| Language pair | Chapters | Sentences (thousands) | Source words (millions) | Target words (millions) |
|---|---|---|---|---|
| en-es | 2,933 (1,022) | 1,060 (80) | 26.6 (6.5) | 27.7 (6.9) |
| en-fr | 2,934 (1,022) | 1,100 (80) | 27.8 (6.5) | 28.8 (7.2) |
| en-de | 2,818 (1,006) | 1,100 (79) | 27.2 (6.4) | 25.1 (6.2) |

**Table 1:** Approximate sizes of the collections used to train and test (in parenthesis) the document translation retrieval system.

The probabilistic language models for the languages involved have been obtained with the SRILM language modelling toolkit (Stolcke, 2002). Additionally, the MOSES open-source decoder (Koehn et al., 2007) provided the phrase-based statistical machine translation system complemented with the open-source tool GIZA++ (Och and Ney, 2003), which provided the word alignments that MOSES requires as initialization. The training phase[2] consists of the following steps:

1. Word-alignments in the training subset are computed using GIZA++.
2. All bilingual phrases consistent with the word alignments (overlapping or not) are extracted and used by MOSES to create the translation model.
3. A 5-gram language model is obtained with the SRILM toolkit from the same training subset.
4. The *minimum error training* algorithm (MERT, Och (2003)) is used by MOSES to optimize the model over the development subset.

After the probabilistic models have been obtained, the test phase proceeds as follows:

1. All chapters in the test set are indexed with the Apache Lucene[3] text search engine (Cafarelle and Cutting, 2004). The stemmers and the default list of stopwords provided by Lucene are used both when indexing the documents and when analyzing the query.
2. The translation of the segments in the source text maximizing Eq. (2) are used to generate the queries for the search engine.
3. The performance of the system is evaluated in terms of the success rate, that is, how often the target document is among the first $D$ documents retrieved. As the *Europarl* corpus contains exactly one target document for each source document, the success rate coincides with the *average recall* (Manning et al., 2008, p. 155).

[2]See `http://www.statmt.org/wmt09/baseline.html` for a detailed description of MOSES training options.

[3]Available at `http://lucene.apache.org`

| Num. docs retrieved | Words in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 1 | 0.02 | 0.05 | 0.13 | 0.27 |
| 2 | 0.04 | 0.08 | 0.20 | 0.35 |
| 5 | 0.07 | 0.12 | 0.28 | 0.46 |
| 10 | 0.10 | 0.17 | 0.35 | 0.54 |
| 20 | 0.13 | 0.22 | 0.43 | 0.62 |

**Table 2:** Baseline success rate obtained with a query made of randomly selected words as a function of the number of documents retrieved and the number of words in the query.

# 4 Results and discussion

Often, the main language of a text can be safely determined: either the document metadata provide this information or simple language models can guess it with a high reliability. Furthermore, word overlap between a query and a (wrong) document is more probable when the document and the query are expressed in the same language. Therefore, in the experiments below, success rates for different strategies are computed with respect to the documents in the target subset $\Omega_T$. Furthermore, as the results showed no significant dependence on $T$, only the average results for the 3 target languages are presented here.

The performance of our approach is compared to that of a baseline which generates a query with a random selection of $W$ words; these words are among those obtained after translating, with the same SMT system, the first 5 sentences in the source document. Search engines allow one to group consecutive words as a single term by surrounding them with quotation marks and, then, the document must match the words exactly in the order they are written in the group. Therefore, we have also obtained the average recall obtained with a fixed number of terms of variable size, which were selected randomly among groups of consecutive words in the five first sentences of the translated document. The experiments showed that the success rate improved only if short terms (no longer than four words) were used, the best results being obtained when they contained about three words. However, this rate was never higher than that obtained with an identical number of words selected randomly: compare for instance the column for 5 words in table 2 with that for 2 terms (6 words) in table 3.

These baseline results have been compared with those obtained with different variations of the statistical approach and summarized in tables 4–6. The first one, table 4, presents the success rate when the query consists of those phrases generated by the SMT system that maximize the quotient between the translation probability of the phrase and its document frequency. As expected, the larger the number of terms (and, therefore, the number of words) in the query and the larger the number of documents retrieved, the

| Num. docs retrieved | Terms in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 1 | 0.06 | 0.11 | 0.25 | 0.43 |
| 2 | 0.09 | 0.15 | 0.33 | 0.53 |
| 5 | 0.13 | 0.21 | 0.42 | 0.64 |
| 10 | 0.16 | 0.27 | 0.49 | 0.71 |
| 20 | 0.20 | 0.32 | 0.57 | 0.77 |

**Table 3:** Baseline success rate obtained with a query made of randomly selected 3-word terms as a function of the number of documents retrieved and the number of 3-word terms in the query.

| Num. docs retrieved | Terms in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 1 | 0.06 | 0.12 | 0.24 | 0.41 |
| 2 | 0.08 | 0.15 | 0.30 | 0.49 |
| 5 | 0.10 | 0.18 | 0.36 | 0.57 |
| 10 | 0.11 | 0.20 | 0.41 | 0.63 |
| 20 | 0.13 | 0.23 | 0.45 | 0.68 |

**Table 4:** Success rate obtained with a query made of the most discriminative phrases generated by the SMT system when translating whole sentences in the source document.

higher the probability that the document translation is among the documents retrieved. However, there is no clear improvement with respect to the results in table 2, even if the average term has 4 words or more. Inspection of the phrases selected indicates that they tend to be long phrases with very few occurrences in the training corpus. On the one hand, long phrases are less frequent than shorter ones and, on the other hand, their translation probabilities tend to be higher due to the scarcity of sentences in the training set containing them. Both factors contribute to increase the quotient in Eq. (2) used to select the most discriminative phrases but also increase the uncertainty in the estimation of the probabilities involved in the maximization.

In order to alleviate the scarcity of training data for the estimation of the document frequency of long phrases, we allowed for the selection of single words as terms. Table 5 shows the success rate when only single words are included in the query; these words are selected because they maximize the quotient between the translation probability of the phrases containing them and the corresponding word document frequencies. As can be seen in this table, the recall figures can be as high as 3 times larger than those obtained with the baseline approach and clearly improve those obtained with the whole-phrase strategy.

However, words in scarce phrases may still have large scores due to the high transla-

| Num. docs retrieved | Words in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 1 | 0.20 | 0.33 | 0.60 | 0.84 |
| 2 | 0.26 | 0.43 | 0.70 | 0.89 |
| 5 | 0.30 | 0.50 | 0.78 | 0.93 |
| 10 | 0.31 | 0.53 | 0.83 | 0.95 |
| 20 | 0.33 | 0.55 | 0.85 | 0.96 |

**Table 5:** Success rate obtained with a query made of the most discriminative words in phrases generated by the SMT system when translating whole sentences in the source document.

| Docs retrieved | Words in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 1 | 0.32 | 0.51 | 0.84 | 0.97 |
| 2 | 0.43 | 0.63 | 0.90 | 0.98 |
| 5 | 0.51 | 0.73 | 0.95 | 0.99 |
| 10 | 0.55 | 0.77 | 0.97 | 1.00 |
| 20 | 0.56 | 0.80 | 0.98 | 1.00 |

**Table 6:** Success rate obtained with a query made of the most discriminative words in phrases generated by the SMT system when translating single words in the source document.

tion probabilities of these phrases. Table 6 shows the success rate when both the translation input and the query terms are single words, and the query consists of those words maximizing the quotient between the word translation probability and the word document frequency. The translation is still provided by the SMT system trained with the same corpus and, therefore, the output can include multi-word expressions as phrases. However, each word in the output is evaluated separately.

The results obtained in this way clearly improve over the previous ones as, using only 5 words, the desired document is among the first 5 documents retrieved with probability $0.95$. This result indicates that the reliable estimation of the probabilities is more important than the plausibility of the probabilistic models.

Finally, in order to check if the approach is stable with respect to the number of indexed documents, we have computed the success rate when the number of indexed documents grows. As can be seen in table 7, the effect of the number of documents is moderate, especially if a significant number of words are included in the query, which makes this approach useful for larger collections.

| Num. docs indexed | Words in query 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| 50 | 0.54 | 0.87 | 0.97 | 1.00 |
| 100 | 0.53 | 0.80 | 0.93 | 0.97 |
| 500 | 0.47 | 0.80 | 0.93 | 1.00 |
| 1,000 | 0.43 | 0.77 | 0.93 | 1.00 |

**Table 7:** Success rate obtained with the best strategy when one document is retrieved as a function of the number of documents indexed.

# 5 Conclusions

We have explored three different strategies, based on statistical machine translation techniques, to select the most effective query terms when looking for the translation of a given document. Among the strategies explored, the best results are obtained when the source document is translated as a sequence of single words and queries consist of those words in the phrases obtained after the translation which maximize the quotient between the probability that the source word is translated into the phrase containing the query word and the document frequency of the query word. The experiments also show that the approach is robust with respect to the collection size. We plan to apply the method to the identification of translated passages in comparable corpora, that is, collections that contain multilingual versions of documents which are only partially parallel (such as Wikipedia).

The source code used for all the experiments (about 1,500 lines written in C++) is distributed under the terms of the GNU General Public License and can be downloaded from `http://code.google.com/p/doctrans`.

## Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

# References

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 222–229, New York, NY, USA.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cafarelle, M. and Cutting, D. (2004). Building nutch: Open source search. *Queue*, 2(2):54–61.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluation the role of BLEU in machine translation research. In *EACL 2006, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Einstein, A. (1911). Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes. *Annalen der Physik*, 340(10):898–908.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 167–174, New York, NY, USA.

Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA.

Hiemstra, D. and de Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In *Research and Advanced Technology for Digital Libraries, Proceedings of the Third European Conference, ECDL'99*, pages 274–293, Paris, France.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information Processing & Management*, 41(3):433–455.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. Available at `http://www.statmt.org/europarl/`.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Demo and Poster Sessions, 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 48–54, Edmonton, AL., Canada.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214, Morristown, NJ, USA.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, USA.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, USA.

Sánchez-Villamil, E., Santos-Antón, S., Ortiz-Rojas, S., and Forcada, M. L. (2006). Evaluation of alignment methods for HTML parallel text. In *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, volume 4139 of *Lecture Notes in Computer Science*, pages 280–290, Turku, Finland. Springer-Verlag.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *KI 2002, Advances in Artificial Intelligence: Proceedings of the 25th Annual German Conference on AI*, volume 2479 of *Lecture Notes in Computer Science*, pages 18–32, Berlin. Springer-Verlag.