Initial approaches on Cross-Lingual Information Retrieval using Statistical Machine Translation on User Queries

Marta R. Costa-jussà, Christian Paz-Trillo and Renata Wassermann

[1] Computer Science Department
Institute of Mathematics and Statistics
University of São Paulo, Brazil
Rua do Matão 1010, São Paulo, SP 05508-090
{martarcj, cpaz, renata}@ime.usp.br

Abstract. *In this paper we propose a multilingual extension for OnAIR which is an ontology-aided information retrieval system applied to retrieve clips from a video collection. The multilingual extension basically involves allowing the user to search in several languages in a multilingual video collection. Particularly, the pair of languages we work in this paper are English and Portuguese. In order to perform query translation we use a statistical machine translation approach. Our experiments show that the multilingual system is capable of achieving almost the same quality of that obtained by the monolingual system.*

Resumo. *Neste trabalho, propomos uma extensão multilingue para OnAir que é um sistema de recuperação de informação auxiliado por uma ontologia. O sistema é usado para recuperar clips de uma coleção de vídeos. A extensão multilingue permite ao usuário fazer buscas em duas línguas em uma coleção de vídeo multilingue. Particularmente, o par de línguas que trabalhamos neste artigo são Inglês e Português. Para realizar a conversão de consulta, usamos uma abordagem estatística de tradução. As nossas experiências mostraram que o sistema multilingue é capaz de atingir quase a mesma qualidade do obtido pelo sistema monolingue.*

1. Introduction

The information society is generating a vast quantity of multilingual information. Recently, there is a growing interest in looking for information in digital videos. Generally, the user can save time, by avoiding to browse through hours of video in order to find the information he is looking for. Additionally, these videos may be in a foreign language. Although he may be able to understand the foreign language, he may not be able to formulate a query. This is the application we are focusing on in this paper in the context of the OnAIR (Ontology-Aided Information Retrieval) system. OnAIR, started in 2003, intended to allow users to look for information in video fragments through queries in natural language. The idea is save the user from the time consuming experience of having to browse through hours of video in order to find an answer for his questions.

The main contribution of this paper is the experimentation of concatenating a state-of-the-art SMT system together with an IR retrieval system that uses ontologies. This concatenation has been done for the Brazilian-Portuguese/English language pair and it can be easily be extended to other pair of languages.

The remaining of this paper is organized as follows. Next section briefly explains the related work in the area of Cross-language Information Retrieval. Section 3 describes the OnAIR structure and architecture. Then, section 4 is dedicated to the OnAIR cross-language extension. Finally, experiments and conclusions are reported in sections 5 and 6, respectively.

2. Related Work

The multilingual extension of OnAIR is basically a challenge of cross-language information retrieval (CLIR). Given a query in a source language, the aim of CLIR is retrieving related documents in a target language. (Oard and Diekema 1998) identified four types of strategies for matching a query with a set of documents in the context of CLIR by: cognate matching, document translation, query translation or interlingua techniques. From these techniques the most used are the query translation and the interlingua techniques.

Query translation methods translate user queries to the language that the documents are written. It is the most popular approach in CLIR experimental systems due to its tractability and convenience. CLIR through query translation methods has been mainly faced by using dictionary-based (i.e. using machine-readable dictionaries, MRD), machine translation (MT) and/or parallel texts techniques (Chen and Bao 2009). Among the different machine translation techniques, we have the corpus-based techniques such as statistical or example-based (Way and Gough 2005) and the rule-based techniques (Forcada 2006). In this paper we are using one of the most popular approaches nowadays which is the standard phrase-based statistical machine translation (SMT) approach (Koehn et al. 2007a).

Interlingua methods translate both documents and queries into a third representation. The approach aims at associating related textual contents among different languages by means of language-independent semantic representations. The conventional interlingua-based CLIR approach uses latent semantic indexing (LSI) for constructing a multilingual vector-space representation of a given parallel document collection (Deerwester et al. 1990; Dumais et al. 1996; Chew and Abdelali 2007). Such a representation is known to be noisy and sparse. That is why in order to obtain more efficient vector-space representations, space reduction techniques such as latent semantic indexing and probabilistic latent semantic indexing (Hofmann 1999) are applied. The new reduced-space dimensions are supposed to capture semantic relations among the words and the documents in the collection. Recent approaches have achieved interesting results by using regression canonical correlation analysis (an extension of canonical correlation analysis) where one of the dimensions is fixed and demonstrate how it can be solved efficiently (Rupnik and Shawe-Taylor 2008).

3. The OnAir system

OnAIR is in essence an information retrieval system which has been described in detail in previous studies such as (Paz-Trillo et al. 2005). In this section we briefly describe the most relevant characteristics of the system. First, we show how the information retrieval is done and, second, we show how a monolingual ontology is used for query expansion.

3.1. Information Retrieval

OnAIR relies on the vector space model (Baeza-Yates and Ribeiro-Neto 1999)for information retrieval. It was built to receive videos and keywords or their transcriptions, with timeline markers, as input, and to allow the users to query for video excerpts using natural language. When a user query is presented, OnAIR returns a list of video excerpts that best answer the user query.

The video transcriptions are pre-processed, using traditional IR techniques: stemming and stopword removal, then the vector space model is used for indexing and retrieving. As usual in traditional IR systems, some additional techniques are needed to avoid natural language difficulties like Polysemy and Synonymy.

3.2. Ontology description

Ontologies are defined in general as an explicit specification for a conceptualization (Gruber 1993). As mainly used for Information Retrieval it can be seen as a set of concepts related by hierarchies and other kind of properties in a specific domain (Ding 2001). Ontologies have been commonly used in IR through query expansion and conceptual distance measures (Paz-Trillo et al. 2005).

A domain ontology related to the topics from the videos is needed to be able to do the query expansion. By definition, query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In particular, the domain ontology is used to measure the conceptual distance among seed query terms and new ones.

4. Cross-lingual extension

In general, a statistical machine translation system relies on the translation of a source language sentence s into a target language sentence $\hat{t}$. Among all possible target language sentences t we choose the one with the highest probability, as show in equation (1):

$$\begin{aligned} \hat{t} &= \arg\max_{t} [P(t|s)] && (1) \\ &= \arg\max_{t} [P(t)\, P(s|t)] && (2) \end{aligned}$$

The probability decomposition shown in equation (2) is based on Bayes' theorem and it is known as the noisy channel approach to statistical machine translation (Brown et al. 1990). It allows to model independently the target language model $P(t)$ and the source translation model $P(s|t)$. The basic idea of this approach is to segment the given source sentence s into segments of one or more words, then each source segment is translated and the target sentence is composed from these segment translations. On the one hand, the translation model weights how likely words in the foreign language are translation of words in the source language; the language model, on the other hand, measures the fluency of hypothesis t. The search process is represented as the arg max operation.

The translation model in the phrase-based approach (Koehn et al. 2003) is composed of phrases. A phrase is a pair of m source words and n target words extracted from a parallel sentence that belongs to a bilingual corpus. The parallel sentences have previously been aligned at the word level (Brown et al. 1993). Then, given a parallel sentence aligned at the word level, phrases are extracted following the next criteria: we consider the words that are consecutive in both source and target sides and which are consistent with the word alignment. We consider a phrase is consistent with the word alignment if no word inside the phrase is aligned with one word outside the phrase. Finally, phrase translation probabilities are estimated as relative frequencies (Zens et al. 2002).

A language model assigns a probability to each target sentence. Standard language models are computed following the n-gram strategy, which considers sequences of n words. In order to compute the probability of an n-gram, it is assumed that the probability of observing the ith word in the context history of the preceding i-1 words can be approximated by the probability of observing it in the shortened context history of the preceding n-1 words. The main problem with this modeling is that it assigns probability zero to strings that have never seen before. One way to solve this problem is assigning non-zero probabilities to sentences they have never seen before by means of smoothing techniques (Kneser and Ney 1995).

A variation of the so-called noisy channel approach is the log-linear model (Och and Ney 2002). It allows using several models or so-called features and to weight them independently as can be seen in equation (3):

$$\hat{t} = \arg\max_{t} \left[\sum_{m=1}^{M} \lambda_m h_m(s,t) \right] \quad (3)$$

This equation should be interpreted as a maximum-entropy framework and as a generalization of equation (2) (Zens et al. 2002).

Most common additional features that are used in the maximum-entropy frameword (in addition to the standard translation and language model) are the lexical models, the word bonus and the reordering model. The lexical models are particularly useful in cases where the translation model may be sparse. For example, for phrases which may have appeared few times the translation model probability may not be well estimated. Then, the lexical models provide a probability among words (Brown et al. 1993) and they can be computed in both directions source-to-target and target-to-source. The word bonus is used to compensate the language model which benefits shorter outputs. The reordering model is used to provide reordering between phrases. For example, the lexicalized reordering model (Tillman 2004) classifies phrases by the movement they made relative to the previous used phrase, i.e., for each phrase the model learns how likely it is followed by the previous phrase (monotonous), swapped with it (swap) or not connected at all (discontinuous).

The different features or models are optimized in the decoder following the minimum error rate procedure (Och 2003). This algorithm searches for weights minimizing a given error measure, or, equivalently, maximizing a given translation metric. This algorithm enables the weights to be optimized so that the decoder produces the best translations (according to some automatic metric and one or more references) on a development set of parallel sentences.

5. Evaluation Framework

This section introduces the details of the evaluation framework. We report the translation and the information retrieval system details including corpus statistics, a description of how we built the systems and the evaluation details.

5.1. SMT data

The parallel corpus used to train the SMT system is taken from the Brazilian-Portuguese-English bilingual collections of the online issue of the scientific news Brazilian magazine REVISTA PESQUISA FAPESP (Aziz and Specia 2011). See statistics in Table 1.

		PT-BR	EN
Train	Sentences	160k	160k
	Words	4,1M	4,3M
	Vocabulary	99,5k	74.7k
Development	Sentences	1375	1375
	Words	34.3k	37.6k
	Vocabulary	6.8k	5.7k
Test	Sentences	1608	1608
	Words	36.8k	38.3k
	Vocabulary	7.3k	6.2k

Table 1. Basic characteristics of the SMT experimental dataset.

5.2. IR data

For testing the information retrieval system in Portuguese-Brazilian we used a video collection compiled from interviews with Ana Teixeira, a Brazilian artist. The interviews were made by Paula P. Braga, the domain expert and there have been used in previous studies as (Paz-Trillo et al. 2005). The interview was developed in the domain of contemporary art and the system uses a domain ontology to expand queries with related terms. To test the system, a battery of queries was synthesized both for English and Brazilian-Portuguese. Statistics of these queries and the corresponding documents for retrieving are shown in Table 2.

		PT-BR	EN
Query	Number	50	50
	Words	349	435
	Vocabulary	155	145
Documents	Number	48	-
	Words	8.2k	-
	Vocabulary	2.4k	-

Table 2. Basic characteristics of the query and documents dataset for the Ana Teixerira videos.

5.3. Translation system

In this paper, we use a system that combines the translation and the language model together with the following additional feature functions: the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. All these features have been described in section 4.

Our translation system was built using MOSES (Koehn et al. 2007b). We used the default MOSES parameters. Word alignment (built with the standard software GIZA++ (Och and Ney 2003)) was performed in both direction source-to-target and target-to-source. These word alignments were merged by using the so-called symmetrization of the grow-diagonal-final-and which is a sophisticated extension of the standard union operation (Koehn et al. 2005). For the translation model, we used phrases up to length 10. Phrase probability is estimated including relative frequencies in both directions (source-to-target and target-to-source), lexical weights and phrase bonus. The lexicalized reordering (Tillman 2004) is used to provide reordering accross sentences. The language model used a 5-gram with Kneser-Ney smoothing. Finally, the word bonus was used to compensate the preference of the language model for shorter outputs. All these different features were combined in equation (3) and the optimization was done using MERT software (Och 2003).

In order to evaluate the translation quality, we used BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2001) which is one of the most popular SMT automatic evaluation metrics. BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. BLEU's output is a number between 0 and 1. This value indicates how similar the candidate translation and reference texts are, with values closer to 1 representing more similar texts.

We evaluated the SMT quality using in-domain and out-domain tests. The former is the one corresponding to the REVISTA PESQUISA FAPESP as shown in Table 1. The out-domain test corresponds to the queries used to test the complete CLIR system as shown in Table 2. Table 3 shows the results in terms of BLEU of the translation system when evaluated in-domain and out-domain.

Test	EN-> PT-BR
In-domain	0.3649
Out-domain	0.1506

Table 3. Evaluation of the translation system in terms of BLEU.

Coherently with international evaluations such as WMT (Callison-Burch et al. 2011), the out-domain test set has a lower performance than the in-domain test set.

5.4. Comparing IR and CLIR system's performance

We performed the following experiments: two experiments using a monolingual information retrieval, recovered from previous publications (Paz-Trillo et al. 2005), and one using a cross-lingual information system. We describe the corresponding systems as follows:

1. IR system: the original system analyzed was the system described in section 3, with two configurations: mono-keywords, which uses only the keywords for retrieval and; mono-kw-fulltext-05 which uses the results of retrieval using keywords and transcriptions, the best configuration for OnAIR as described in (Paz-Trillo et al. 2005)
2. CLIR system (smt-kw-fulltext-05): this system is the concatenation of the statistical machine translation system described in the previous section and the information retrieval system from the point above in this list.

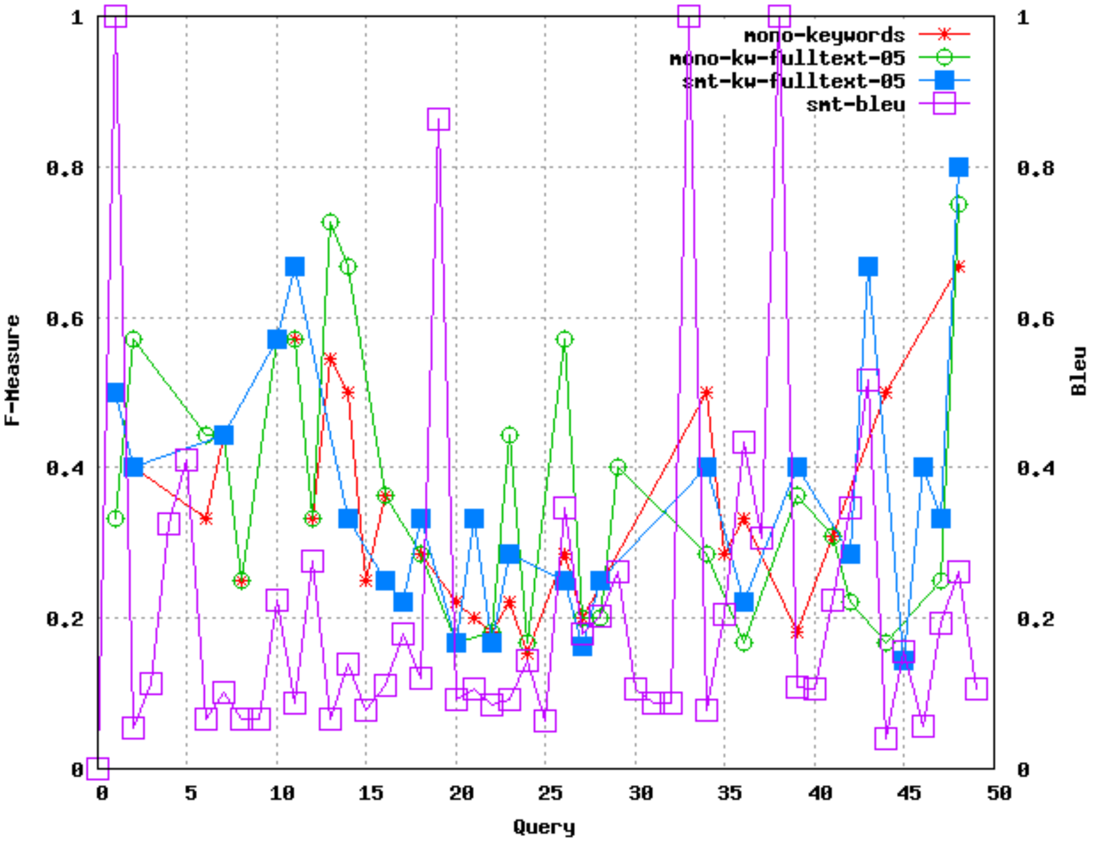

Figure 1. F-measure for the systems analyzed.

Figure 1 shows the results of the f-measure run over the 50 queries analyzed in our experiments in the three configurations presented above and the BLEU measure for the translation of each query.

Surprisingly, experiments show that the CLIR system, for specific queries, is capable of outperforming the IR system. For these queries, the translation system uses a more adequate word, which means that it would be possible to use machine translation to perform query expansion. It would be interesting to built the CLIR system with the n-best translations.

Figure 2 shows the f-measure in average for all systems that we experimented. Here, we observe that the f-measure of with respect to the CLIR system (smt-kw-fulltext-05) is slightly worst than its comparable IR system (mono-kw-fulltext-05). However, in

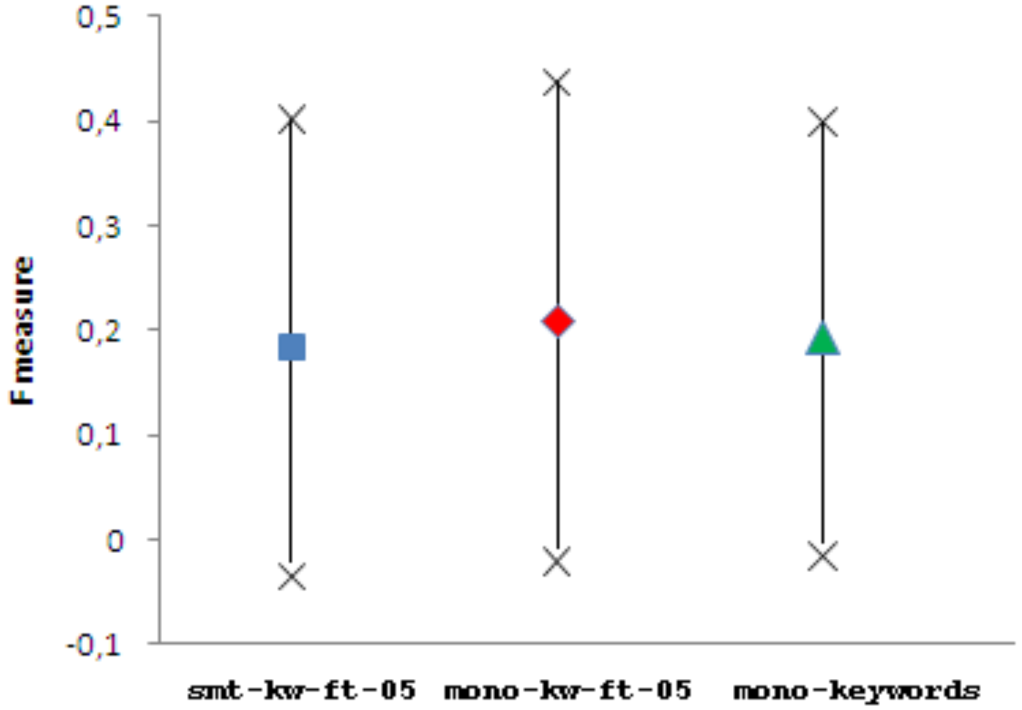

Figure 2. Average f-measure for the systems analyzed.

average, the f-measure using SMT is not highly affected when compared to the best monolingual result.

Finally, Figure 3 shows some translation examples. It shows the input to the CLIR system (smt-kw-fulltext-05), the corresponding translation and the corresponding reference (i.e. the input of the IR system). The two first examples report cases where the CLIR system performs worse than the IR system (mono-kw-fulltext-05) in terms of f-measure. The second two examples report cases where the CLIR system performs better than the IR system in terms of f-measure. Coherently, in the first case, the translation shows a poorer quality than in the second case.

6. Conclusions and future work

This paper has shown an ongoing work that generates a cross-lingual extension for the OnAIR system, which is in essence an information retrieval system using ontologies to expand queries. The cross-lingual extension has been done using a state-of-the-art statistical machine translation system. Experiments show that the best configuration for the IR system uses the results of retrieval using keywords and transcriptions. For the CLIR system, we can get competitive results using a state-of-the-art statistical machine translation system.

As further work, we want to explore different linguistic and statistical techniques (focusing on morphology and semantics) to be introduced in the state-of-the-art statistical MT system in order to correctly translate queries which are out-of-domain of the training corpus. Also it would be interesting to use MT as a query expansion method.

INPUT: How did you become an artist?
TRANSLATION: Como o senhor se um artista?
REFERENCE: Como você virou artista

INPUT: Do you make only interventions or also paintings, sculpture, etc?
TRANSLATION: O senhor faz apenas intervenções ou também pinturas, escultura etc?
REFERENCE: Você só faz intervenções ou faz também pintura, escultura, etc?

INPUT: I loved his work.
TRANSLATION: Adorei seu trabalho.
REFERENCE: Adorei seu trabalho.

INPUT: Have you ever exposed abroad?
TRANSLATION: O senhor já exposta no exterior?
REFERENCE: Você já expôs no exterior?

Figure 3. Translation examples.

7. Acknowledgements

This work has been supported by FAPESP through the OnAir project (2010/19111-9) and the visiting researcher program (2012/02131-2), and by the Spanish Ministry of Economy and Competitiveness through the BUCEADOR project (TEC2009-14094-C04-01) and the Juan de la Cierva fellowship program.

References

[Aziz and Specia 2011] Aziz, W. and Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In STIL 2011, Cuiabá, MT.

[Baeza-Yates and Ribeiro-Neto 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley Longman.

[Brown et al. 1990] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., 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 et al. 1993] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311.

[Callison-Burch et al. 2011] Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22–64, Edinburgh, Scotland.

[Chen and Bao 2009] Chen, J. and Bao, Y. (2009). Cross-language search: The case of google language tools. First Monday, 14(3-2).

[Chew and Abdelali 2007] Chew, P. and Abdelali, A. (2007). Benefits of the passively parallel rosetta stone? Cross-Language information retrieval with over 30 languages. In Proc of the 45th Annual Meeting of the Association for Computational Linguistics, volume 45, page 872.

[Deerwester et al. 1990] Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.

[Ding 2001] Ding, Y. (2001). Ir and ai: The role of ontology. In International Conference of Asian Digital Libraries.

[Dumais et al. 1996] Dumais, S. T., Landauer, T. K., and Littman, M. L. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In SIGIR96 Workshop on Cross-Linguistic Information Retrieval.

[Forcada 2006] Forcada, M. L. (2006). Open-source machine translation: an opportunity for minor languages. In Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).

[Gruber 1993] Gruber, T. R. (1993). A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199–220.

[Hofmann 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proceedings of Uncertainty in Artificial Intelligence, UAI99, pages 289–296.

[Kneser and Ney 1995] Kneser, R. and Ney, H. (1995). Improved backing-off for n-gram language modeling. In IEEE Inte. Conf. on Acoustics, Speech and Signal Processing, pages 49–52, Detroit, MI.

[Koehn et al. 2005] Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In Proceedings of the Int. Workshop on Spoken Language Translation (IWSLT'05), Pittsburg, USA.

[Koehn et al. 2007a] 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. (2007a). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), pages 177–180, Prague, Czech Republic.

[Koehn et al. 2007b] 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. (2007b). Moses: Open source toolkit for statistical machine translation. In Proc. of the ACL, pages 177–180, Prague, Czech Republic.

[Koehn et al. 2003] Koehn, P., Och, F., and Marcu, D. (2003). Statistical Phrase-Based Translation. In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics.

[Oard and Diekema 1998] Oard, D. W. and Diekema, A. R. (1998). Cross-Language information retrieval. Annual Review of Information Science and Technology (ARIST), 33:223–256.

[Och 2003] Och, F. (2003). Minimum Error Rate Training In Statistical Machine Translation. In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics, pages 160–167.

[Och and Ney 2002] Och, F. and Ney, H. (2002). Dicriminative training and maximum entropy models for statistical machine translation. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pages 295–302, Philadelphia, PA.

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

[Papineni et al. 2001] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. IBM Research Report, RC22176.

[Paz-Trillo et al. 2005] Paz-Trillo, C., Wassermann, R., and Braga, P. P. (2005). An information retrieval application using ontologies. J. Braz. Comp. Soc., 11(2):17–31.

[Rupnik and Shawe-Taylor 2008] Rupnik, J. and Shawe-Taylor, J. (2008). Multi-view canonical correlation analysis and cross-lingual information retrieval. In http://videolectures.net/lms08_rupnik_rcca/.

[Tillman 2004] Tillman, C. (2004). A Block Orientation Model for Statistical Machine Translation. In HLT-NAACL.

[Way and Gough 2005] Way, A. and Gough, N. (2005). Comparing example-based and statistical machine translation. Natural Language Engineering, 11(3):295–309.

[Zens et al. 2002] Zens, R., Och, F., and Ney, H. (2002). Phrase-based statistical machine translation. In Verlag, S., editor, Proc. German Conference on Artificial Intelligence (KI).