Using Cross-Language Information Retrieval for Machine Translation

Dhouha Bouamor, Christophe Servan, Nasredine Semmar, Ali Jaoua

▶ **To cite this version:**

Dhouha Bouamor, Christophe Servan, Nasredine Semmar, Ali Jaoua. Using Cross-Language Information Retrieval for Machine Translation. 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Nov 2011, Poznań, Poland. hal-01490031

HAL Id: hal-01490031
https://hal.science/hal-01490031

Submitted on 6 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using Cross-Language Information Retrieval for Machine Translation

Nasredine Semmar[1], Christophe Servan[1], Dhouha Bouamor[1], Ali Jaoua[2]

[1] CEA, LIST, Vision and Content Engineering Laboratory
18 route du Panorama, Fontenay-aux-Roses, F-92265, France
{nasredine.semmar; christophe.servan; dhouha.bouamor}@cea.fr

[2] Computer Science and Engineering Department
University of Qatar, Doha, Qatar
jaoua@qu.edu.qa

Abstract

This paper describes a new machine translation approach based on a statistical language model and a cross-language search engine. This approach consists in building a database of sentences in the target language and considering each sentence to translate as a "query" to that database. Linguistic information such as lemmas, part-of-speech and syntactic dependency relations corresponding to the words of the sentences returned by the cross-language search engine are then combined with a statistical model of the target language to produce a correct translation. This approach has been evaluated by using the Europarl parallel corpus and the BLEU score. The experimental results we obtained are encouraging and demonstrate the effectiveness of the proposed approach.

Keywords: cross-language information retrieval, machine translation, bilingual lexicon, language model, finite-state machine, conditional random field

1. Introduction

Non-availability of parallel corpora, morphology and syntactic structure differences between source and target languages are the major challenges for the use of statistical machine translation. The main idea of our machine translation approach is to use only a monolingual corpus in the target language. This corpus is analyzed syntactically and stored in the database of a cross-language search engine (Grefenstette, 1997). The sentence to translate is considered as a query to this search engine. The cross-language search engine returns a set of sentences in the target language with their linguistic information (lemma, grammatical category, gender, number and syntactic dependency relations). These data are used with a statistical language model learned from the target language corpus to find the correct translations.

The remainder of this paper is organized as follows. We present in section 2 some related work. Section 3 describes the theoretical principles of our machine translation approach and the main components of our cross-language search engine. We discuss in section 4 translation results obtained after submitting 1000 English sentences as queries to the search engine database composed of 1 million French sentences. Section 5 concludes our study and presents our future work.

2. Related work

There are mainly two approaches for machine translation: rule-based and corpus-based (Trujillo, 1999) (Hutchins, 2005). The first and oldest rule-based strategy is the direct translation approach which considers that the translation is direct form the source text to the target text without syntactic or semantic analysis. The second rule-based strategy is the transfer approach. This approach operates in three stages: The first stage converts source language texts into syntactic trees; the second stage converts these trees into their equivalent in target language and the third stage generates the translation. The third rule-based strategy is the interlingua approach which uses an intermediate semantico-syntactic representation which allows generation into any target language without needing to provide specific source-target conversion rules. The corpus-based machine translation approaches use statistics and probability calculations in order to identify equivalences between texts in the corpus. The first corpus-based strategy is the example-based approach which consists of reusing examples of already existing translations as the basis for new translations. This approach operates in three stages: The first stage looks for examples of the corpus which are similar with the sentence to translate. The second stage uses an alignment procedure to identify which parts of the corresponding translation are to be reused. The third stage checks if the reusable parts in example identified during alignment are assembled correctly. The second corpus-based strategy is the statistical approach (Koehn, 2010) which consists in searching for a target language string that maximizes the probability that this string is the translation of a source target string (translation model) and the probability that this target language string is a valid sentence (language model). This approach uses frequency of co-occurrence of strings in aligned texts in order to build the translation model and succession of strings (bigrams and trigrams) in order to build the language model. Rule-based approaches require manual development of bilingual lexicons and linguistic rules, which can be costly, and which often do not generalize to other languages. Corpus-based approaches are effective only when large amounts of parallel text corpora are available. Hybrid approaches combine the strengths of rule-based and corpus-based machine translation strategies (Somers, 2005). (Koehn et al. 2010) reported that within the framework of factored and tree-based translation models, additional linguistic information (lemma, part-of-speech and morphological information) can be successfully exploited to overcome some short-comings of the currently dominant phrase based statistical

machine translation approach and produce promising results.

Our machine translation approach uses linguistic information as in factored translation models but in opposition to these models it does not use parallel corpora. It just needs monolingual corpora in the target language.

3. Machine Translation based on Cross-language Information Retrieval

The machine translation process of our approach is composed of three steps (Fig. 1):

- Constitution of a referring language model for the target language;
- Providing syntactic structures of the translation candidates;
- Generating the correct translation.

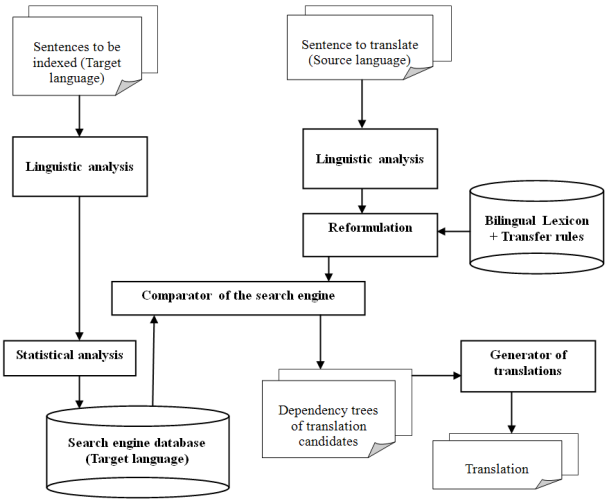

Fig. 1: Machine translation using Cross-language information retrieval

3.1. Constitution of a referring language model for the target language

This step consists in crawling the Web in order to get the maximum texts in the target language, applying a linguistic analysis on these texts and storing the results in a textual database. Each sentence of these texts is considered as a document in our cross-language search engine. For our experimentation, we analyzed and indexed 1 million French sentences of the Europarl parallel corpus[1].

3.2. Providing syntactic structures of the translation candidates

This step uses our cross-language search engine (Semmar et al., 2006) to provide a collection of sentences in the target language. These sentences are considered are translation candidates. Our search engine uses a weighted Boolean model, in which sentences are grouped into classes characterized by the same set of concepts composed of words. The classes constitute a discrete partition of the textual database. This search engine uses a deep linguistic analysis for the query (sentence to translate) and for the indexed sentences, and is composed of the following modules (Fig. 1):

- A multilingual analyzer (LIMA) (Besançon et al., 2010) which includes a morphological analyzer, a part-of-speech tagger and a syntactic analyzer. The linguistic analyzer processes both sentences to be indexed and queries to produce a set of normalized lemmas, a set of named entities and a set of nominal compounds with their morpho-syntactic tags. The syntactic analyzer is used to split sentence words into nominal and verbal chains and recognize dependency relations by using a set of hand written syntactic rules. A set of dependency relations to link nouns to other nouns, a noun with a proper noun, a proper noun with the post nominal adjective and a noun with a post nominal adjective are implemented.
- A statistical analyzer, which computes for target language sentences concept weights based on concept database frequencies. The weight is maximum for words appearing in one single sentence and minimum for words appearing in all the sentences. Inverted files of these sentences are then created and stored in a database.
- A reformulator, to translate the words and to transform the syntactic structure of the sentence to translate into the target language. This reformulator uses an English-French bilingual lexicon composed of 220 000 entries to translate words, and a set of rules to transform syntactic structures from the source language to the target language.
- A comparator, which computes intersections between words and the syntactic structure of the sentence to translate and words and syntactic structures of the indexed sentences. This comparator provides a relevance weight for each intersection and returns the translation candidates. These translation candidates could be sub-sentences composed of only some words corresponding to the translation of just a part of the sentence to translate. Linguistic information such as lemmas, grammatical categories, gender, number and syntactic dependency relations are associated with the words of the translation candidates.

3.3. Generating the correct translation

Automatic generation is the process which consists in producing automatically a natural language text. It uses resources which are not necessarily linguistic. This process is issued from the first translation systems. Automatic generation is a full part of Natural Language Processing (Chomsky, 1956) (Yngve, 1961) (Friedman, 1971) (Melčuk, 1988). It is used in several research domains such as Question/Answering, Automatic summarization, etc. In machine translation, this process is called “text synthesis” as opposed to text understanding or analysis process. Analysis process consists in producing a linguistic structure from texts. The text synthesis process starts from linguistic structures to produce texts.

1 The Europarl parallel corpus is available on http://www.statmt.org/europarl.

Generating the correct translation of our approach consists, on the one hand, in composing the sub-sentences returned by the comparator of the cross-language search engine in order to build a dependency syntactic structure in the target language which covers the sentence to translate, and, on the other hand, in producing a correct sentence in the target language by using the syntactic structure of the translation candidate.

The generation process is composed of two modules: a reformulator and a flexor. The reformulator uses the parts of sentences to match the translation hypothesis. Some linguistic rules are used to assemble the new hypothesis in a lattice of translations. This lattice contains linguistic information for each word of the translation. A statistical model is learned on a monolingual lemmatized corpus which contains linguistic information. This model scores the lattice in order to find the best syntactic hypothesis in the target language. The lattice is implemented by using the AT&T FSM toolkit (Mohri et al., 2002). The language model is learned with the CRF++ toolkit (Kudo and Matsumoto, 2001). The flexor transforms the lemmas of the target language sentence into plain words. We use the linguistic information returned by the cross-language search engine to produce the right form of the lemma. This flexor consists in transforming the lemma of a word into the surface form of this word by using the grammatical category, the gender and the number of the word. For example, the lemma "*avoir*" (verb) in present simple and third person singular will be transformed into the form "*a*". Sometimes, we obtain several forms for the same lemma. To disambiguate, we use a statistical language model based on CRF that has been previously trained on a monolingual corpus. This disambiguation provides the right flexion of the lemma and therefore the best translation (Fig. 2).

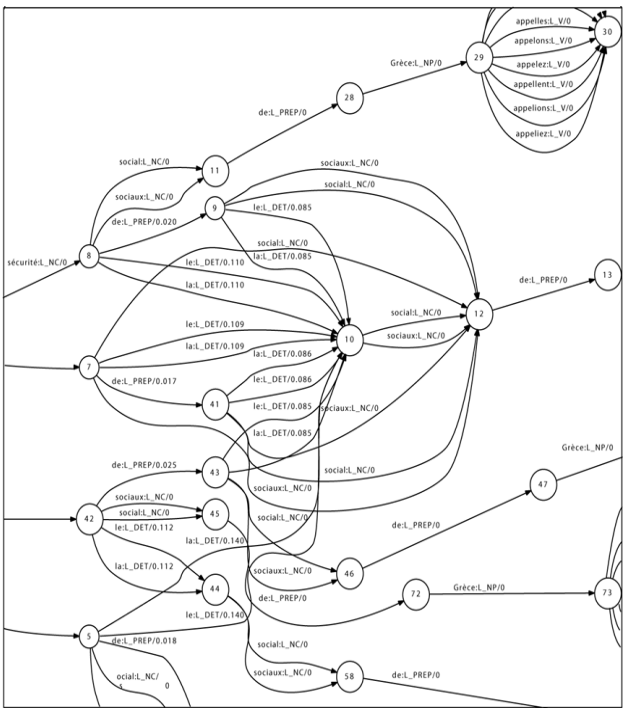

Fig. 2: Results of the flexor for the translation candidates

When linguistic data are too few (for example, missing of the tense for a verb) the flexor produces a set of variations. We enrich our lattice of hypothesis with flexion hypothesis. The whole lattice is scored with another language model, learned from texts in target language.

4. Experiment Results and Discussion

To evaluate the performance of our machine translation approach, we indexed 1 million French sentences of the Europarl corpus and we used a subset of Arcade-II[2] corpus composed of 1000 sentences in English and French. These English sentences which are aligned to their French counterparts are considered as the translation reference. Our translation approach obtained a BLEU score of 31.33%. This score is satisfactory taking into account that only 1 million sentences are indexed and used to train the language model.

Table 1 shows the translation results ordered by their relevance given by our machine translation approach for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*".

Relevance	Translation candidate
1	les fonds de la sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux
2	les fonds de sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux
3	les fonds de la sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des fonds
4	les fonds de sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des fonds
5	les fonds de le sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des capitaux
6	les fonds de le sécurité sociale en Grèce appellent à l'autonomie concernant l'investissement des fonds
7	les fonds de la sécurité social en Grèce appellent à l'autonomie concernant l'investissement des capitaux
8	les fonds de la sécurité social en Grèce appellent à l'autonomie concernant l'investissement des fonds
9	les fonds de le sécurité social en Grèce appellent à l'autonomie concernant l'investissement des capitaux
10	les fonds de le sécurité social en Grèce appellent à l'autonomie concernant l'investissement des fonds

Table 1. The translation results for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*"

2 The ARCADE-II parallel corpus was produced within the French national project ARCADE-II (Evaluation of Sentence and word alignment tools).

In order to analyze the behavior of our machine translation approach on a sentence which has a translation in the cross-language search engine database, we submitted the English sentence "*The Report provides an overview of the health status of Canadians.*". Despite the fact that this sentence exists in the Europarl indexed corpus, our machine translation approach returns other correct translations which are different from the translation which exists in the search engine database (Table 2).

Proposed translation	Reference
la rapport prévoit une panorama de la situation la santé des canadiens.	Dans le Rapport, on donne un aperçu de l'état de santé de la population canadienne.

Table 2. The translation result and the translation reference for the English sentence "*The Report provides an overview of the health status of Canadians.*"

Analysis of the translation results shows that some errors remain. The origins of these errors are different: morpho-syntactic analysis, language model, etc. For example, the English word "*report*" was identified by the morpho-syntactic analyzer as a noun in singular without a specific gender. Consequently, having the French definite article "*la*" before the word "*rapport*" is grammatically correct. The same remark is valid for the English word "*overview*". On the other hand, the English expression "*the health status*" is translated as "*la situation la santé*" instead of "*la situation de la santé*". This is due to the fact that the English expression does not contain the preposition "*of*".

5. Conclusion

We presented in this paper a new approach for machine translation. This approach is based on cross-language information retrieval and needs only monolingual texts in the target language. The first results of our experiments are satisfactory and promising despite the fact that the indexed corpus is small and does not cover all the aspects of the target language. We expect that indexing a large corpus could improve significantly the BLEU score. Analysis of these results showed that we can improve the translation quality by combining a statistical language model with an efficient morpho-syntactic analyzer. In future work, we plan, on the one hand, to perform a large scale evaluation of our approach by using the CESTA[3] evaluation package, and, on the other hand, to adapt it for new languages pairs such as English-Arabic and French-Arabic.

Acknowledgment

This research work is supported by the FINANCIALWATCH (QNRF NPRP: 08-583-1-101) project.

3 The CESTA evaluation package was produced within the French national project CESTA (Evaluation of MT systems), as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT).

References

Besançon, R., De Chalendar, G., Ferret, O., Gara, F., Laib, M., Mesnard, O., and Semmar, N. (2010). *LIMA: A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation*. *The seventh international conference on Language Resources and Evaluation, Valetta, Malta, May 19-21, 2010.*

Chomsky, N. (1958). *Three models for the description of language*. IRE Transactions on Information Theory.

Friedman, J. (1971). *A computer model of transformational grammar*. Elsevier.

Grefenstette, G. (1999). *Cross-language information retrieval*. Boston: Kluwer Academic Publishers.

Hutchins, J. (2005). *Machine Translation: General Overview*. Oxford University Press.

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

Koehn, P. and Hoang H. (2007). *Factored Translation Models. The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.*

Kudo, T. and Matsumoto Y. (2001). *Chunking with support vector machines. Meeting of the North American chapter of the Association for Computational Linguistics (NAACL).*

Melčuk, I. (1988). . *Dependency Syntax: Theory and Practice*. State University of New York Press.

Mohri, M., Pereira, F. and Riley, M. (2002). Factored Translation Models Weighted Finite-State Transducers in Speech Recognition. In:. *Computer Speech and Language*, 16(1):69-88.

Semmar, N., Laib, M., and Fluhr, C. (2006). *A Deep Linguistic Analysis for Cross-language Information Retrieval. The fifth international conference on Language Resources and Evaluation, Genoa, Italy, May 22-28, 2006.*

Somers, H. (2005). *Machine Translation: Latest Development*. Oxford University Press.

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation*. Springer-Verlag Series on Applied Computing.

Yngve, V. (1996). *Random generation of English sentences. International Conference on Machine Translation of Languages and Applied Languages Analysis.*