A New Hybrid Machine Translation Approach Using Cross-Language Information Retrieval and Only Target Text Corpora

Nasredine Semmar
CEA, LIST, Vision and Content Engineering Laboratory
18 route du Panorama
Fontenay-aux-Roses, F-92265, France
nasredine.semmar@cea.fr

Dhouha Bouamor
CEA, LIST, Vision and Content Engineering Laboratory
18 route du Panorama
Fontenay-aux-Roses, F-92265, France
dhouha.bouamor@cea.fr

Abstract

Parallel corpora play a vital role in Statistical Machine Translation. Non-availability of these corpora is a major barrier for adding new languages pairs. In this paper, we propose a new hybrid approach for English-French machine translation combining a cross-language search engine and a statistical language model trained from a monolingual corpus. The cross-language search engine returns the translation candidates ordered by their relevance and the language model of the target language is used to disambiguate the translation. This approach has been evaluated and compared to Moses. We used 100000 French sentences of the Europarl corpus to train the language model, 1103 English-French sentences of the Arcade-II corpus as the translation reference and the BLEU score. The obtained scores are 21.33% for our approach and 21.45% for Moses. The experimental results also showed that our approach provides better translation performance in terms of grammatical coherence.

1 Introduction

Parallel corpora play a vital role for training translation models in Statistical Machine Translation (SMT). Non-availability of these corpora, morphology and syntactic structure differences between source and target languages are the major challenges for adding new languages pairs for SMT engines. We present, in this paper, a new hybrid approach for machine translation which uses only a monolingual corpus in the target language. This approach is based on a cross-language search engine which returns for each sentence to translate a set of translation candidates extracted from the monolingual corpus already indexed. A statistical language model is then used to identify the correct translation.

This paper is organized as follows. In section 2, some related work is presented. Section 3 describes the implementation of our hybrid machine translation approach. In section 4, some experimental results are reported and discussed. Section 5 concludes our study and presents our future work.

2 Related Work

There are two main approaches for machine translation (Trujillo, 1999) (Hutchins, 2005):

- Rule-based approaches.
- Corpus-based approaches.

The rule-based approaches regroup word-to-word translation, syntactic translation with transfer rules and interlingua which uses an intermediate semantico-syntactic representation to generate translations into any target language.

The corpus-based machine translation approaches use statistics and probability calculation in order to identify equivalences between texts in the corpus (Koehn, 2010). This probability calculation depends on two measures. The first is the probability that the words in the target language are translations of the words in the source language (translation model). The second is the probability that these words are correctly combined in the target language (language model). Probability that a given word in the target text is a translation of a given word in the source text is calculated on the basis of a sentence-aligned parallel corpus. The language model consists of probabilities of sequences of words based on a monolingual corpus in the target language.

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) presented an extension of the state-of-the-art phrase-based statistical machine translation models in order to integrate additional linguistic information such as lemmas, part-of-speech, and morphological properties of words. The authors reported that experiments showed gains over standard phrase-based models, both in terms of automatic scores (gains of up to 2% BLEU), as well as a measure of grammatical coherence.

Our hybrid approach for machine translation is based on a new paradigm which consists in using a cross-language search engine to extract translated texts from a monolingual corpus and combining linguistic information with a statistical language model in order to generate the correct translation.

3 Machine Translation Based on Cross-language Information Retrieval

Cross-language information retrieval consists in providing a query in one language and searching documents in different languages (Grefenstette, 1997), and the goal of machine translation is to produce for each sentence in the source language its equivalent in the target language. Cross-language information retrieval using linguistic analysis for indexing and interrogation and rule-based machine translation are closely related domains. Both use bilingual lexicons and automatic text analysis.

The machine translation prototype implementing our approach is composed of two modules: A cross-language search engine and a text generator (Figure 1):

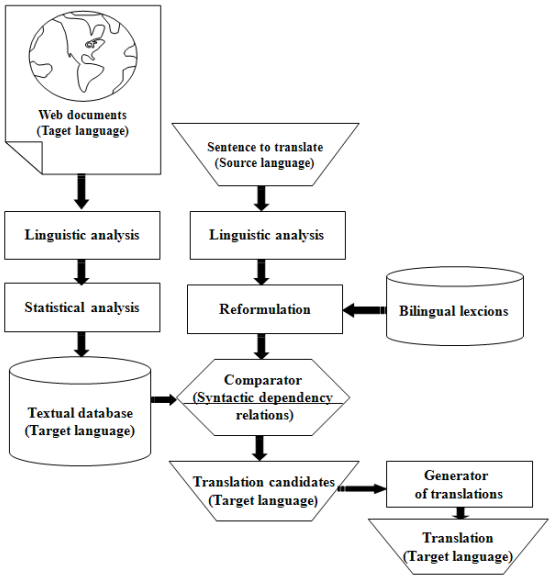

Figure 1: Machine translation using cross-language information retrieval

3.1 Cross-language Information Retrieval

The cross-language search engine (Semmar et al., 2006) is used to provide a collection of sentences in the target language. These sentences are considered are translation candidates. The search engine uses a weighted Boolean model, in which sentences in the target language are grouped into classes characterized by the same set of concepts composed of words. This search engine is composed of a multilingual analyzer, a statistical analyzer, a reformulator and a comparator.

Multilingual Analysis

The multilingual analysis is built using a traditional architecture (LIMA) (Besançon et al., 2010) and includes a morphological analyzer, a part-of-speech tagger and a syntactic analyzer. The linguistic analyzer produces a set of normalized

lemmas, a set of named entities and a set of dependency relations between words.

Statistical Analysis

The role of the statistical analysis is to attribute to each word or a compound word a weight according to the information it provides to choose the target sentences relevant to the sentence to translate. The weight is maximum for words appearing in one single sentence and minimum for words appearing in all the sentences. This weight is used by the comparator to compare intersection between the sentence to translate and indexed sentences. Our search engine uses a weighted Boolean model, in which sentences are grouped into classes characterized by the same set of concepts. The classes constitute a discrete partition of the database. For example, if the sentence to translate is "*nuclear waste*" on a database containing only sentences on nuclear plants, the statistical model indicates that sentences containing the compound word "*nuclear waste*" are more relevant than sentences containing the words "*nuclear*" and "*waste*". Sentences containing the words "*nuclear*" and "*waste*" are more relevant than sentences containing only the word "*waste*".

Query Reformulation

Reformulation consists in inferring new words from the original query (sentence to translate) words according to lexical and semantic knowledge (synonyms, etc.). The reformulation can be used to increase the quality of the retrieval in a monolingual interrogation (Debili, 1989). It can also be used to infer words in other languages. The query terms are translated using bilingual dictionaries. Each term of the query is translated into several terms in target language. The translated words form the search terms of the reformulated query. The links between the search terms and the query concepts can also be weighted by a confidence value indicating the relevance of the translation. Reformulation can be achieved on the word or on the word with a specific part of speech and can also be used to transform the syntactic structure of the sentence to translate into the target language. This reformulation uses an English-French bilingual lexicon composed of 220000 entries to translate words, and a set of rules to transform syntactic structures from the source language to the target language.

Comparison of the sentence to translate with indexed sentences

The comparator 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.

3.2 Text Generation

Our text generation approach is based on a syntactic analysis. This 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 text generator 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.

4 Experiment Results and Discussion

To evaluate the performance of our machine translation approach, we indexed the first 100000 French sentences of the Europarl[1] corpus and we used a subset of Arcade-II[2] corpus composed of 1103 sentences in English and French as the translation reference. In order to compare the translation results of our approach with the results of the open source baseline system Moses, we used the same Europal bilingual corpus composed of the first 100000 sentences in English and French to train the language and translation models and we considered the same 1103 sentences of Arcade-II as a test corpus. We also considered that there is only one reference per test sentence and we used the BLEU score to evaluate the translation quality of the two systems. Our translation approach obtained a score of 21.33% and Moses obtained a score of 21.45%. These two scores are very close and are satisfactory taking into account that only 100000 sentences are used to train these two systems.

In order to show the relevance of using a deep linguistic analysis in machine translation, we used Google Translate [3] to translate into French the sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*". Google Translate proposes the translation "*Administrations de sécurité sociale en Grèce sont appelant à l'indépendance à l'égard de l'investissement de capitaux.*". Thereby, the compound word "*Social security funds*" has been translated by the compound word "*Administrations de sécurité sociale*" and the expression "*are calling for*" has been translated as "*sont appelant*".

As we can see, our translation prototype proposes the compound word "*fonds de la sécurité sociale*" as a translation for the compound word "*Social security funds*" and the expression "*appellent à*" as a translation for the expression "*are calling for*". These translations are better than those provided by Google Translate.

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.

Table 1: The first five translations returned for the English sentence "*Social security funds in Greece are calling for independence with regard to the investment of capital.*"

5 Conclusion and Future Work

This paper proposed a new hybrid approach for English-French machine translation combining a cross-language search engine and a statistical language model trained from a monolingual

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

[2] The Arcade-II parallel corpus was produced within the French national project Arcade-II (Evaluation of sentence and word alignment tools), as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT).

[3] This experimentation has been done in March 2011. At present, Google Translate proposes a better translation.

corpus. The results we obtained showed that it is possible to improve machine translation performance by combining a good bilingual lexicon with a large statistical language model. In addition, using a deep linguistic analysis on the sentence to translate and also on the indexed sentences allowed the search engine to present relevant translations on the top of the list of the translation candidates. In order to confirm these results, we are currently working on a large evaluation of our approach and in the same time we are adapting it for a new language pair English-Arabic.

Acknowledgments

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

References

Besançon R., De Chalendar G., Ferret O., Gara F., Laib M., Mesnard O., and Semmar N. 2010. A Deep Linguistic Analysis for Cross-language Information Retrieval LIMA :A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. Proceedings of LREC 2010.

Debili, F., Fluhr, C., and Radasoa, P. 1989. About reformulation in full text IRS. Information Processing & Management, Infortmation Processing and Mmanagement, Elsevier.

Grefenstette G. 1999. Cross-language information retrieval. Boston: Kluwer Academic Publishers.

Hutchins J. 2005. Machine Translation: General Overview. The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, UK.

Koehn P. 2010. Statistical Machine Translation. Cambridge University Press.

Koehn P., Haddow B., Williams P., and Hoang H. 2010. More Linguistic Annotation for Statistical Machine Translation. Proceedings of the Fifth Workshop on Statistical Machine Translation and MetricsMATR.

Kudo T. and Matsumoto Y. 2001. Chunking with support vector machines. Meeting of the North American chapter of the Association for Computational Linguistics (NAACL), 1–8.

Mohri M., Pereira, F., and Riley M. 2002. Factored Translation Models Weighted Finite-State Transducers in Speech Recognition. 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. Proceedings of LREC 2006.

Somers H. 2005. Machine Translation: Latest Developments. The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, UK.

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