A Hybrid Approach for Machine Translation Based on Cross-language Information Retrieval

Christophe Servan, Nasredine Semmar

A Hybrid Approach for Machine Translation Based on Cross-language Information Retrieval

Christophe Servan, Nasredine Semmar

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

Abstract

This paper presents a hybrid approach for Machine Translation (MT) based on Cross-language Information Retrieval (CLIR). This approach uses linguistic and statistical processing and does not need parallel corpora as linguistic resources. A first experimental evaluation of this approach has been done on the CESTA corpus and the obtained results seem good and encouraging. The next step is the TALK evaluation of the IWSLT2010 Workshop.

1. Introduction

Parallel corpora sources are only available for a limited number of language pairs and the process of building these corpora is time consuming and expensive.

The main idea behind the CEA-LIST machine translation prototype is to use only mono-lingual corpora. These corpora can be collected from the Web. First, we make a syntactic analysis of the target language corpus. The result is given as database to our search engine. Then, the sentence to translate is considered as a query to our search engine. The search engine returns a set of sentences with their linguistic information. We use this information to translate the sentence associated to a monolingual model learned from the target language corpus. We associate linguistic information and the statistical model to translate the source language sentence.

This paper is structured as the following. In section 2, we present some related work. Sections 3 and 4 describe the theoretical concepts and the implementation of our machine translation prototype. Section 5 presents some experimental results and section 6 concludes our study and presents our future work.

2. Related work

The Web gives us access to a vast amount of information in many languages. The knowledge of these languages is a way to access to information. Recently, the research in automatic or semi-automatic translation increased due to the commercial demand.

There many approaches for machine translation, but the three main ones are:

- rule-based approaches which use linguistic resources such as lexicons and syntactic rules [1];
- statistical approaches based on IBM models [2];
- example-based approaches [3].

These approaches can be combined in order to produce better translations. They are called hybrid approaches [4].

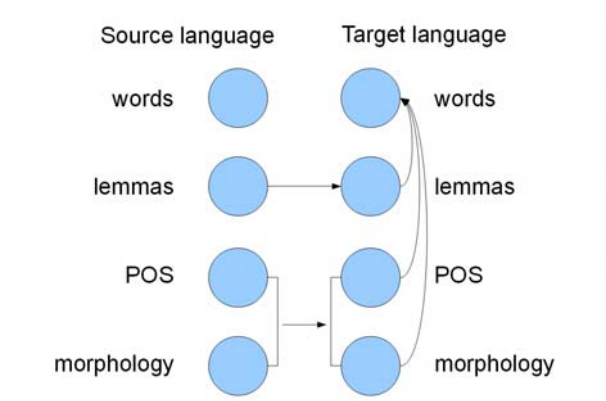

Figure 1: Factored model that inspires our approach

In our approach, we use a new paradigm of statistical machine translation: the factored translation model (fig. 1). This model uses the words and some associated linguistic information [5]. But in our approach, we do not use parallel corpora (or bitext). We propose an approach which needs only mono-lingual corpora that we can find easily on the Internet.

This approach uses the CEA-LIST cross-language search engine to extract translated texts from the monolo-lingual corpus. Then, we merge our formal translation approach with statistical models. The next section describes in details our machine translation prototype.

3. The CEA LIST machine translation approach

Our approach aims to be independent from bitexts. In this way, to create translations, our system gets through three important steps:

- extraction;
- reformulation;
- flexion.

Theses parts use some tools developed at the CEA-LIST LVIC laboratory.

3.1. Extraction

The first step extracts a collection of sentences which contains parts of the source language sentence candidates for translation. This step is mainly a part of the CEA-LIST cross-language search engine [6]. It returns a set of sentence parts

that contains translation hypothesis. The target language translations hypothesis extracted are lemmas.

We use a set of rules to extract linguistic information associated to theses parts. The linguistic information are lemmas, part-of-speech and syntactic relations. As presented before, we use all theses linguistic data in the next steps.

3.2. Generation

Automatic generation is the process which consists to produce automatically a natural language text. It uses resources that are not necessary linguistic. This process is issued from the first translation systems. Automatic generation is a full part of Natural Language Processing. It is used in several research domains such as Question/Answering, Automatic summarization etc.

In machine translation, we rather call this process "text synthesis" opposed to text understanding or analysis process. Analysis process consists to produce a linguistic structure from text. The text synthesis process starts from linguistic structures to produce text.

3.2.1. State of the Art

Text Generation evolved through three phases. Firstly, theses systems were based on the Chomsky generative grammar [7]. These systems are made to validate syntactic theories. As example, we can mention Yngre [8] and Friedman [9] systems.

Then, another system family used semantic data only. They were limited to the current sentence. The dialog aspect was not taking in account. Theses systems can not organize a set of ideas. Most of translation systems using this approach are based on Igor Melčuk approach [10]. The ETAP-3 generator system [11] uses this approach.

At last, generators uses systemic grammar approaches like KPML [12] or SURGE [13]. These grammars use trees as syntactic representation, especially, dependency grammar. Generation process contains a linear process to transform trees into linear sentences [14].

Recent systems use systematic and pragmatic aspect of text. But they are generally used in systems where the unit is not the sentence but a set of sentences, like in automatic summarization.

As text generators in machine translation systems, we can mention RealPro [15] and AlethGen [16].

3.2.2. The CEA-LIST Generation approach

Our approach is based on a syntactic analysis. This analysis gives us some linguistics data to generate the target language structure. The result is a set of hypothesis.

This structure is enriched with translation hypothesis given by the search engine. These translations are lemma that we have to transform in plain words.

The flexion operation transforms the set of translation hypothesis which is scored to obtain the n-best translation. The next section details our system implementation.

4. Implementation

Our Generator is composed of two steps. The first one gives us a hypothesis of syntactic structure of the target language sentence; the second one gives us the flexion hypothesis. The figure 2 shows the entire translation process.

We show an example of translation step-by-step. The source language sentence is "Social security funds in Greece are calling for independence with regard to the investment of capital."

4.1. Reformulation

This step uses the parts of sentences to match the translation hypothesis. We use some linguistic rules to assemble our new hypothesis in a lattice of translations. This lattice contains linguistics data for each translation. We also enrich the lattice with syntactic rules. Theses rules create syntactic translations from the source language to the target language. For example, the translation from English to French creates some linguistic permutations like adjectives associated to a noun:

- English: "*the green mountain*";
- French: "*la montagne verte*".

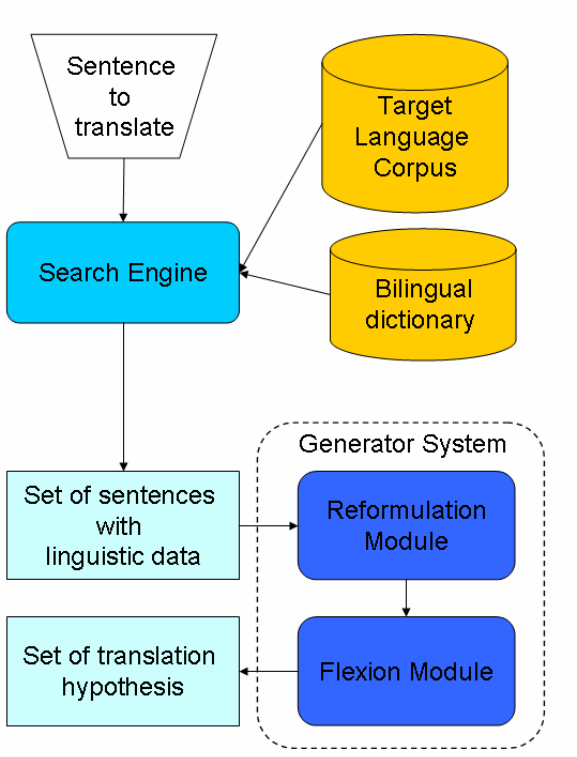

Figure 2 : *The CEA-LIST Machine Translation prototype architecture*

The last part of this step uses a statistical model. This model is learned on a mono lingual lemmatized corpus which contains linguistic data. This language model scores our lattice in order to give the best syntactic hypothesis in the target language.

In order to implement our lattice, we use the AT&T FSM toolkit [17]. The language model is learned with the CRF++

toolkit [18]. The choice of using the Conditional Random Fields for modeling our language model is related to the use of the left and the right sentence context.

Figure 3 show the reformation result applied to our example: "Social security funds in Greece are calling for independence with regard to the investment of capital."

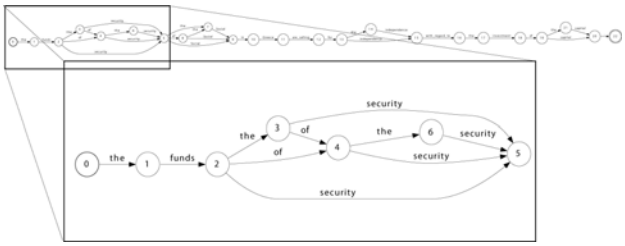

Figure 3: Example of syntactic reformulation

4.2. Flexion

The last part of our system transforms the lemmas of the target language sentence into plain words. We use the linguistic data to give the right form of the lemma. This new form is given by the LIMA (CEA-LIST Multilingual Analyzer) flexion server [19].

Table 1: Translation results given by our system for our example

N-Best	Translation hypothesis
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

When linguistic data are too few, (i.e. the missing of the tense for a verb) the flexion server gives 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. We use the same toolkit of the reformulation step. The result is a set of translation hypothesis of the source language sentences (Table 1).

5. Experimental results

The evaluation process of our machine translation prototype with the whole data of the CESTA campaign is on progress. However, we did a preliminary evaluation on a small set of sentences and the translation results are very encouraging.

For example, the following table illustrates the translation results for sentence "The report provides an overview of the health status of Canadians."

Table 2: Translation results given by our system for the sentence "The Report provides an overview of the health status of Canadians."

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.

Analysis of the translation results shows that some errors remain. The origins of these errors are different: errors of the morpho-syntactic analyzer, link words too many, error from the 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 contains only one preposition "of".

6. Conclusion

We presented in this paper the CEA-LIST Machine Translation tool. This tool is based on a cross-language information retrieval approach. The first results of our experiments seem good and promising. Analysis of these results showed that there is still room for improving the translation quality by using in particular a more efficient morpho-syntactic analyzer.

In future work, we plan do consolidate our results by evaluating this Machine Translation prototype on a large set of examples and to adapt it for new languages pairs.

7. Acknowledgements

This research work is supported by the ANR WEBCROSSLING project (ANR - Programme Technologies Logicielles - 2007).

8. References

[1] A. Trujillo, “Translation Engines: Techniques for Machine Translation”, Applied Computing, Springer, 1999.

[2] P. Koehn, “Statistical Machine Translation”, Cambridge University Press, 2010.

[3] D. Wu, "MT model space: statistical versus compositional versus example-based machine translation", Machine Translation 19 (3-4), 2005.

[4] H. Schwenk, D. Déchelotte, H. Bonneau-Maynard and A. Allauzen, "Modèles statistiques enrichis par la syntaxe pour la traduction automatique", TALN 2007, Toulouse, 2007.

[5] P. Koehn and H. Hoang, “Factored Translation Models”, EMNLP-2007, 2007.

[6] N. Semmar, M. Laib, and C. Fluhr, “A Deep Linguistic Analysis for Cross-language Information Retrieval”, LREC 2006, Italy, 2006.

[7] N. Chomsky, “Three models for the description of language”, IRE Transactions on Information Theory, 1956.

[8] V. Yngve, “Random generation of English sentences”, International Conference on Machine Translation of Languages and Applied Languages Analysis, 1961.

[9] J. Friedman, “A computer model of transformational grammar”, Elsevier, 1971.

[10] I. Melčuk, “Dependency Syntax : Theory and Practice”, State Univ. of New York Press, 1988.

[11] Ju. D. Apresjan, I.M. Boguslavskij, L.L. Iomdin, A.V. Lazurskij, V.Z. Sannikov et L. L. Tsinman, “Système de traduction automatique {ETAP}”, La Traductique P.Bouillon and A.Clas (eds). Montreal, Les Presses de l'Universite de Montreal, 1993.

[12] J. A. Bateman, “Enabling technology for multilingual natural language generation: the KPML development environment”, Natural Language Engineering, 1997.

[13] M. Elhadad, J. Robin, “SURGE : a comprehensive plug-in syntactic realization component for text generation”, Computational Linguistics, 1999.

[14] J. Vergne, “Between dependency tree and linear order, two transforming processes”, COLING-98, 1998.

[15] B. Lavoie, O. Rambow, “RealPro–a fast, portable sentence realizer“, ANLP’97, 1997.

[16] J. Coch, "Overview of AlethGen", 8th International workshop on Natural Language Generation, Herstmonceux, United Kingdom, 1996.

[17] M. Mohri, F. Pereira, and M. Riley, “Weighted Finite-State Transducers in Speech Recognition”, Computer Speech and Language, 16(1):69-88, 2002.

[18] T. Kudo, and Y. Matsumoto, “Chunking with support vector machines”, Meeting of the North American chapter of the Association for Computational Linguistics (NAACL), pp. 1–8, Pitsburgh, PA, USA, June 2001.

[19] R. Besançon, G. de Chalendar, O. Ferret, F. Gara, M. Laïb, O. Mesnard, and N. Semmar. "LIMA: A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation", LREC-2010, Malte, 2010.

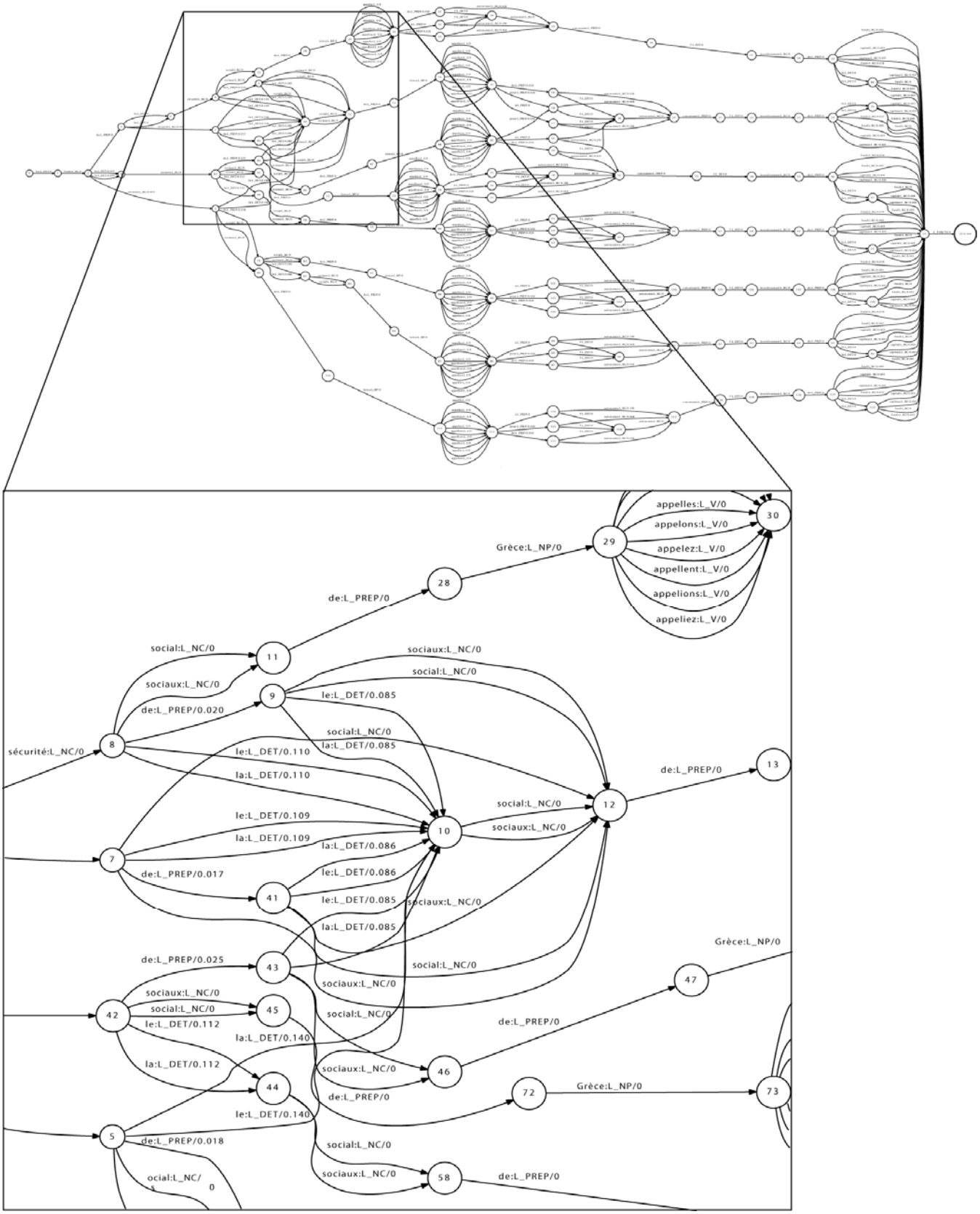

Figure 4 : Translation hypothesis produced by the flexion system