NERITS - A Machine Translation Mashup System Using Wikimeta and DBpedia

Kamel Nebhi, Luka Nerima, and Eric Wehrli

LATL, Department of linguistics
University of Geneva
Switzerland
firstname.name@unige.ch

Abstract. Recently, Machine Translation (MT) has become a quite popular technology in everyday use through Web services such as Google Translate. Although the different MT approaches provide good results, none of them exploits contextual information like Named Entity (NE) to help user comprehension.

In this paper, we present NERITS, a machine translation mashup system using semantic annotation from Wikimeta and Linked Open Data (LOD) provided by DBpedia. The goal of the application is to propose a cross-lingual translation by providing detailed information extracted from DBpedia about persons, locations and organizations in the mother tongue of the user. This helps at scaling the traditional multilingual task of machine translation to cross-lingual applications.

Keywords: Mashup, Machine Translation, Named Entity Recognition, Linked Open Data.

1 Motivation

Machine Translation (MT) has become increasingly commonplace in everyday use through Web services such as BabelFish, Bing Translator or Google Translate. Although the different MT approaches provide good results, none of them exploits contextual information like named entity - such as the names of persons, organizations or locations - to help user comprehension. In addition, communication across languages and cultures has become vitally important, especially now with the globalization of Internet. In this context, cross-lingual knowledge bases like DBpedia can be used to connect structured information across languages. This helps at scaling the traditional multilingual task of machine translation to cross-lingual applications.

In this paper, we present NERITS, a machine translation system using semantic annotation provided by Wikimeta and LOD [5] from DBpedia. Initially, the mashup application translates a sentence using our MT Web service ITS-2[1](Interactive Translation System). Then, we use the Wikimeta API[2] in

[1] http://latlapps.unige.ch/Translate
[2] http://www.wikimeta.com/

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7955, pp. 312–318, 2013.

order to extract named entities in the source sentence and to establish a link to the DBpedia databank. Finally, we give detailed information about the named entities using the DBpedia triplestore.

The goal of NERITS is to propose a cross-lingual translation by providing a set of information about persons, locations and organizations in the mother tongue of the user.

This article is structured as follows: section 2 describes our MT system ITS-2; section 3 provides details on the proposed mashup approach. We conclude and give some perspectives in section 4.

2 ITS-2 : An Interactive Translation System

2.1 Overview

ITS-2 is a large-scale translation system developed in our laboratory, LATL, in the last couple of years [11,13]. The language pairs supported are: English-French, German-French, Italian-French, Spanish-French, French-German and French-English.

At the software level, an object-oriented design has been used, similar to the design adopted for the Fips multilingual parser on which it relies [12]. To a large extent, ITS-2 can be viewed as an extension of the parser. It relies heavily on the detailed linguistic analysis provided by the parser for the supported languages, and exploits the lexical information of its monolingual lexicons. Both systems aim to set up a generic module which can be further refined to suit the specific needs of, respectively, a particular language or a particular language-pair.

The system is based on the familiar transfer architecture, with its three main components, parser, transfer and generation. First, the input sentence is parsed, producing an information-rich phrase-structure representation with associated predicate-argument representations. The parser also identifies multi-word expressions such as idioms and collocations [8] – crucial elements for a translation system .

Then, the transfer module maps the source-language abstract representation into the target-language representation. Given the abstract nature of this level of representation, the mapping operation is relatively simple and can be sketched as follows: recursively traverse the source-language phrase structure in the order: head, right sub-constituents, left subconstituents. Lexical transfer (the mapping of a source-language lexical item to an equivalent target-language item) occurs at the head-transfer level (provided the head is not empty); it yields a target-language equivalent term, often (but by no means always) of the same category. Following the projection principle used in the parser, the target-language structure is projected on the basis of the lexical item which is its head.

However, the projections (i.e., constituents) which have been analyzed as arguments of a predicate undergo a slightly different transfer process, since their precise target-language properties may be in part determined by the subcategorization features of the target-language predicate. To take a simple example, the direct object of the French verb *regarder* in (1-a) will be transferred to English

as a prepositional phrase headed by the preposition *at*, as illustrated in (2-a). This information comes from the lexical database. More specifically, the French-English bilingual lexicon specifies a correspondence between the French lexeme [$_{VP}$ regarder NP] and the English lexeme [$_{VP}$ look [$_{PP}$ at NP]]. For both sentences, we also illustrate the syntactic structures as built by the parser and/or the generator of ITS-2:

(1) a. Paul regardait la voiture.
b. [$_{TP}$ [$_{DP}$ Paul] regardait$_i$ [$_{VP}$ $\mathbf{e}_i$ [$_{DP}$ la [$_{NP}$ voiture]]]]

(2) a. Paul was looking at the car.
b. [$_{TP}$ [$_{DP}$ Paul] was [$_{VP}$ looking [$_{PP}$ at [$_{DP}$ the [$_{NP}$ car]]]]]

2.2 Evaluation

The last evaluations of ITS-2 were for the Fifth and Seventh Workshop on Statistical Machine Translation [2,3]. The LATL participated in the French-English and English-French tasks. The table 1 shows the best results obtained by ITS-2 in terms of BLEU[3] [7] and Translation Edit Rate[4] (TER) [10] using the newstest2010 and newstest2012 corpus as evaluation set.

Table 1. Translation results from French to English and English to French measured on newstest2010 and newstest2012

Pair of language	BLEU	TER
French-English	16.5	0.785
English-French	21.8	0.684

3 The Proposed Approach

NERITS (Named Entity Recognition for the Interactive Translation System) is a Web application plugged on top of various tools such as ITS-2 machine translation, Wikimeta semantic annotation platform and DBpedia triplestore.

3.1 Architecture

For our mashup, we used a layered architecture that includes: data retrieval using ITS-2 and Wikimeta Web services, data integration using DBpedia knowledge base, and the user interface.

[3] BLEU is currently the standard in MT evaluation. It calculates n-gram precision and a brevity penalty, and can make use of multiple reference translations as a way of capturing some of the allowable variation in translation.

[4] TER calculates the number of edits required to change a hypothesis translation into a reference translation. The possible edits in TER include insertion, deletion, and substitution of single words, and an edit which moves sequences of contiguous words.

Actually, the mashup application provides French to English and English to French translation. The figure 1 shows the architecture of our system.

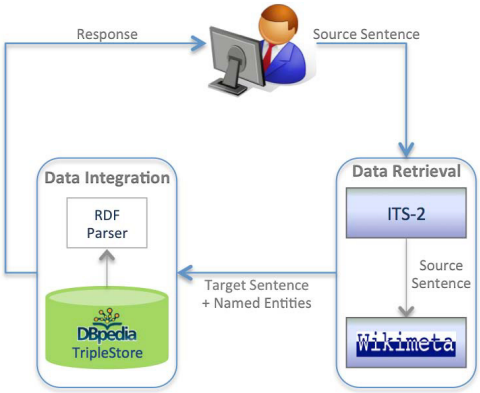


Fig. 1. *NERITS* Mashup Architecture

Data Retrieval using Web Services. The following Web services were used to build the mashup:

1. **ITS-2**: our ReST Web service [6] generates a translation using the machine translation ITS-2 (cf. section 2);
2. **Wikimeta** [4]: the source sentence is analyzed by the semantic annotation platform which performs the traditional task of named entity recognition and also the task of named entity linking[5] [9]. Wikimeta proposes an approach based on a resource of contextual words called *Linked Data Interface* (LDI). The LDI associates several metadata[6] to each entity described in Wikipedia. The disambiguation task uses the *Semantic Disambiguation Algorithm* (SDA), which identifies the item in the LDI that is most similar to the context (the context is represented by the set of words that appear around the NE) of the named entity. The system shows promising results with 90 percent of recall for French and 86 percent of recall in English. The listing 1.1 shows the Wikimeta categorization for the NE "Jean-Marc Ayrault".

```
<extraction>
 <NE>Jean-Marc Ayrault</NE>
 <type>PERS.HUM</type>
 <LOD>http://www.dbpedia.org/resource/Jean-Marc_Ayrault</LOD>
</extraction>
```

Listing 1.1. Wikimeta categorization for the NE "Jean-Marc Ayrault"

[5] Entity linking is the task to link the entity mention in text with the corresponding entity in the existing knowledge bases like DBpedia or GeoNames.

[6] A set of surface forms, a set of words that are contained in the entity description and an URI that points to some entity in the LOD Cloud.

Data Integration. NERITS integrates data by using Semantic Web technology, in particular LOD as DBpedia. DBpedia [1] defines LOD URIs for millions of concepts by extracting structured information from Wikipedia.
The DBpedia databank contains:

- Data about persons, locations, organizations, music albums, etc.
- 4 million datasets described by 1,2 billions of triples
- 7 million of external RDF links to Freebase, GeoNames, YAGO, etc.

The listing 1.2 is a part of the RDF/XML representation for the DBpedia entry "Jean-Marc Ayrault". These data provide several informations about "Jean-Marc Ayrault" such as his birth date, his place of birth, his occupation, etc.

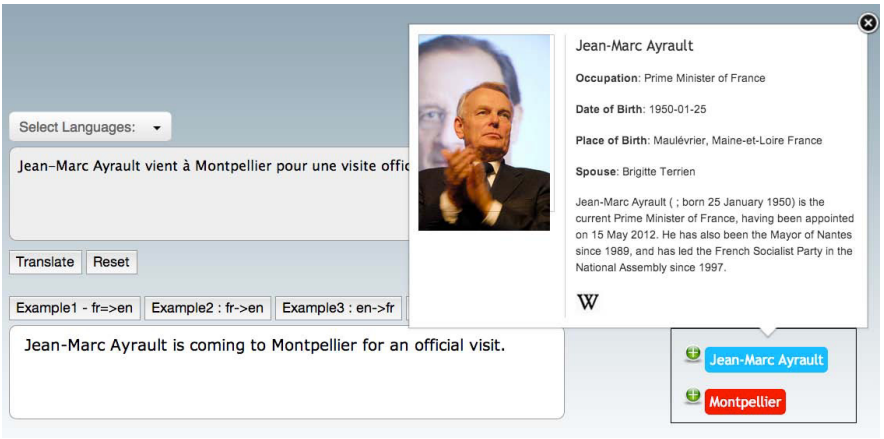
```
<rdf>
 <rdfs:label>Jean-Marc Ayrault</rdfs:label>
 <dbpedia-owl:orderInOffice>Prime Minister of France</dbpedia
     -owl:orderInOffice>
 <dbpprop:birthDate>1950-01-25
 </dbpprop:birthDate>
 <dbpprop:placeOfBirth>Maulevrier, Maine-et-Loire France</
     dbpprop:placeOfBirth>
 <dbpprop:spouse>Brigitte Terrien</dbpprop:spouse>
 <foaf:depiction rdf:resource="http://upload.wikimedia.org/
     wikipedia/commons/b/be/Jean-Marc_Ayrault_-_mars_2012.jpg
     " />
</rdf>
```

Listing 1.2. Part of the RDF/XML representation for the DBpedia entry "Jean-Marc Ayrault"

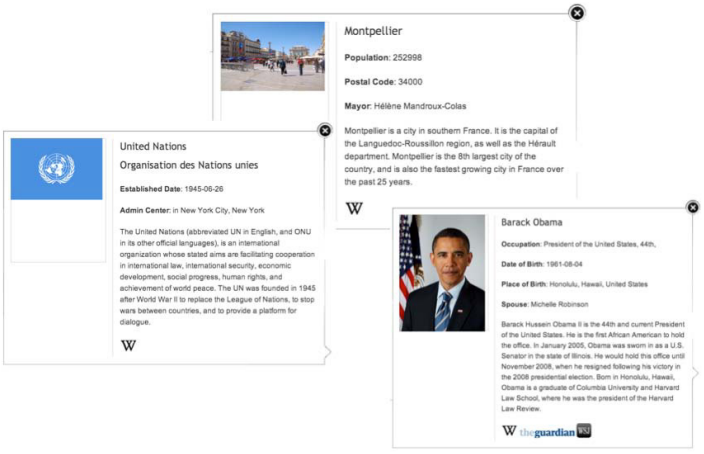
The processing of the data integration module relies on two main steps : when the Wikimeta API recognizes and annotates entities in the source sentence, a query is issued to the DBpedia Triplestore; then a PHP module parses and converts the RDF file into HTML. For our mashup, informations extracted from DBpedia are available in English and French.

User Interface. The user interface is freely available at `http://cms.unige.ch/lettres/linguistique/nebhi/nerits/`.

The figure 2a is a screenshot of the user interface for the translated sentence "Jean-Marc Ayrault vient à Montpellier pour une visite officielle". In a first step, NERITS provides the translation of the source sentence. Then, if the source sentence contains NE, the mashup application provides detailed information about persons, locations and organizations. For the example "Jean-Marc Ayrault", it gives the following information: occupation, date of birth, place of birth, spouse, a brief biography and a link to Wikipedia. For the city "Montpellier", the figure 2b shows the following information: population, country, mayor, a brief description and a link to Wikipedia.



(a) User Interface



(b) Other Examples

Fig. 2. *NERITS* screenshots

The figure 2b also shows other examples provided by NERITS. For the organization "UN", it gives information such as the established date, the location of the admin center, a brief description and a link to Wikipedia. For the example "Barack Obama", in addition to conventional information, it also provides links to a set of newspaper articles from the Guardian and/or the Wall Street Journal.

4 Conclusion - Further Work

In this paper, we have presented a machine translation mashup system using semantic annotation provided by Wikimeta and Linked Open Data. We have given a first glance on how semantic annotation and information from DBpedia can be combined to help user comprehension.

In future work, we plan to incorporate recognition of events (such as *September 11 attacks* or *Fall of the Berlin Wall*). We'll also try to make better use of Linked Open Data in order to provide the points of interest of a Location (for example "Tour Eiffel" in Paris) or people around a Person (for example "Hillary Clinton" is in Barack Obama's entourage). Finally, we'll add other machine translation systems (such as Google Translate, Bing Translator and Babelfish) and languages as German-French and French-German.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web (7), 154–165 (2009)
2. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, pp. 17–53. ACL (2010)
3. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 Workshop on Statistical Machine Translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, Montréal, Canada, pp. 10–51. ACL (2012)
4. Charton, E., Gagnon, M., Ozell, B.: Automatic semantic web annotation of named entities. In: Canadian Conference on AI, pp. 74–85 (2011)
5. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)
6. Nebhi, K.: A ReSTFul Web Service for multilingual LRT. In: 3rd International Conference on the Future of Information Sciences (INFuture): Information Sciences and e-Society, Zagreb, Croatia (2011)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)
8. Seretan, V., Wehrli, E.: Accurate collocation extraction using a multilingual parser. In: ACL (2006)
9. Shen, W., Wang, J., Luo, P., Wang, M.: Linden: linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 449–458. ACM (2012)
10. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas (2006)
11. Wehrli, E.: Translating Idioms. In: Proceedings of COLING 1998, Montreal, pp. 1388–1392 (1998)
12. Wehrli, E.: Fips, a deep linguistic multilingual parser. In: ACL 2007 Workshop on Deep Linguistic Processing, pp. 120–127. Czech Republic, Prague (2007)
13. Wehrli, E., Nerima, L., Scherrer, Y.: Deep linguistic multilingual translation and bilingual dictionaries. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 90–94. ACL (2009)