PROJECT GROUP EFNILEX
Minutes live meeting 31 August 2009, Budapest

Enclosure: Handout for the meeting

Participants:

František Čermák
Aranka Laczházi
John Simpson
Beatrix Tölgyesi
Johan Van Hoorde
Tamás Váradi
Jolanta Zabartskaite
Enikő Héja (minutes)

Absent: Annemieke Hoorntje, Johan Van Hoorde

1. Automatic generation of prototype bilingual dictionaries

In accordance with the conclusions of the previous meeting, Enikő Héja made a report presenting the results of two automatically generated prototype dictionaries. However, since the relevant tools and resources were not at our disposal for the Dutch-French language pair, we have chosen Hungarian-Slovenian, instead of the originally planned Dutch-French language pair. The method will be applicable to Dutch and French as well, once the resources are at our disposal. As originally decided, Hungarian and Lithuanian was the other studied language-pair.

Following the decisions of the previous meeting, we have continued with the method of statistical word alignment on the basis of sentence-aligned parallel corpora. The workflow consists of two main stages: first we have to collect a large amount of general domain parallel texts (cc. 10 million tokens per language). The second task is to build two prototype dictionaries while testing the methodology.

It was also pointed out that instead of the originally planned 20.000-45.000 entries, it is sufficient to include 15.000-25.000 entries in a medium sized dictionary.

(a) Collecting data

As it was envisaged, the main bottleneck of the method is the availability of relevant texts of appropriate size. To tackle this task we have contacted *Rūta Marcinkevičienė* and *Andrius Utka* at the Centre of Computational Linguistics, Vytautas Magnus University. The Centre of Computational Linguistics is supplying the relevant texts and also performs the necessary linguistic annotation.

We have also contacted Tomaz Erjavec at Dept. of Knowledge Technologies, Jozef Stefan Institute who also provided us with a list of texts from the FIDA corpus and drew our attention to a freely available tool, which supplies the needed linguistic annotation of the texts.

We were advised to gather texts from FIDA+ by contacting Simon Krek, the coordinator of the corpus compilation, due to copyright issues.

For both language pairs, the enlargement of the corpora is an ongoing, high-priority process.

Should the coverage of the resulting dictionaries not be sufficient, combining statistical word-alignment with the hub-and-spoke model might be a possible solution for augmenting the coverage of the dictionaries.

(b) Building prototype dictionaries

After creating the parallel corpora, the dictionary building process follows, which is composed of three main steps. The first one is the language dependent pre-processing so that we could find the right headwords to be included in our dictionaries. The next step is the statistical word-alignment, which determines the most probable translation pairs through assigning translation probabilities to word pairs. Finally, we build the dictionaries by associating the relevant sentences to the suitable translation pairs. Since the method itself can be modified based on several parameters, finding the optimal setting requires a delicate evaluation.

(c) Evaluation

The method was evaluated based on two test runs in the case of the Hungarian-Slovenian language pair. Because the size of the dictionary is inversely proportional to the dictionary's accuracy, we have to take into consideration the lexicographers' needs when setting these parameters. In addition, being absolutely based on context, these dictionaries might offer a different approach to the question of what is to be considered as a translation unit than traditional lexicography, which might entail the reinterpretation of notions like 'homonymy' and 'polysemy'. Thus, test versions of dictionaries have to be extensively evaluated for both language pairs. That is why evaluation is the second high priority task for the next phase. Accordingly, after the evaluation of a sample dictionary, Beatrix Tölgyesi and Aranka Laczházi will decide if they could contribute to the evaluation of the Hungarian-Lithuanian prototype.

The main drawback of the word alignment approach is that it cannot handle multiword expressions, so without the extension of this methodology, collocations cannot be included in the dictionaries automatically. One possible solution is that after the previous detection of multiword expressions we treat them as single tokens (this might be done by treating frequent bi-, tri- or even tetragrams separately).

Equally, human lexicographers can deal with this problem, since the provided example sentences should make clear the cases of multiword expressions.

For evaluative purposes the use of a commercial softwer – Paraconc – was suggested.

2. Administrative and financial aspects

- The processing of Lithuaninan texts was done with the help of the Centre of Computational Linguistics, Vytautas Magnus University. They agreed to collaborate with us as sub-contractors to the project and Tamás Váradi, in agreement with Taalunie, has offered to allocate budget resources up to 3000 Euros in remuneration to the Centre.

- The validation of Hungarian-Lithuanian pairings will be undertaken by Beatrix Tölgyesi and Aranka Laczházi or their colleagues out of the budget of the Hungarian team.

- The costs of a Hungarian-Lithuanian pilot dictionary will be assessed on the basis of the validation.

- Estimates for the budget should be prepared before the annual conference.