NLP applications: completing the puzzle

Ruben Izquierdo

Vrije University of Amsterdam, Amsterdam, The Netherlands,
ruben.izquierdobevia@vu.nl

1 Description

The Natural Language and Computational Linguistics communities have traditionally faced different problems with specific approaches and mostly in an isolated manner or in a pipeline way. The former approaches focus on solving one particular aspect of the Natural Language Processing without considering other problems, very easily ending up in incoherent solutions. Pipeline approaches tackle one problem at a time in a sequence of sub-problems, where the output of one step is the input of the next step. These methods suffer from error propagation, tend to be too deterministic (one decision can not be changed later) and lead to sub–optimal solutions. To exemplify this problem, we include the Figure 1, where we show the result of an error analysis that we performed on the participant outputs from SensEval-2 to SemEval-2013. The table in that figure shows the error rate on the monosemous words, which were due mainly to part-of-speech errors (the tagger marks the word as adjective but it is a noun), or errors in the multiword detection (the systems tag "stuck" when they should tag "get_stuck"). In SemEval2010 this error rate reaches 98%. More details on this error analysis can be found here [1].

Competition	Monosemous	Wrong	Examples
Senseval2	499 (20.9%)	37.5%	gene.n (*suppressor_gene.n*), chance.a (*chance.n*) next.r (*next.a*)
Senseval3	334 (16.6%)	44.1%	Datum.n (*data.n*) making.n (*make.v*) out_of_sight (*sight*)
Semeval2007	25 (5.5%)	11.1%	get_stuck.v, lack.v, write_about.v
Semeval2010	31 (2.2%)	97.9%	Tidal_zone.n pine_marten.n roe_deer.n cordgrass.n
Semeval2013 (lemmas)	348 (21.1%)	1.9%	Private_enterprise, developing_country, narrow_margin

Fig. 1. Error rate on monosemous instances

Another aspect that seems to be not fully considered is the role of the context. For example, WSD systems usually restrict the context of a word to a very

narrow window of tokens around the target word, usually not bigger than the sentence in which the token occurs. This is clearly not enough in some cases where the clues for getting the proper meaning of the word are to be found in another part of the document or even outside of this document (background information). Following another example from the error analysis mentioned in the previous paragraph, we include here a comparison of the average performance of the systems on the cases where the most frequent sense applies, and in the rest of cases. Results can be seen in Figure 2, where clearly the systems perform very well on the most frequent cases but this performance drops dramatically in the rest of cases. One reason could be that the systems are not modelling properly the context and they are inducing just to apply the most frequent sense in all the cases.

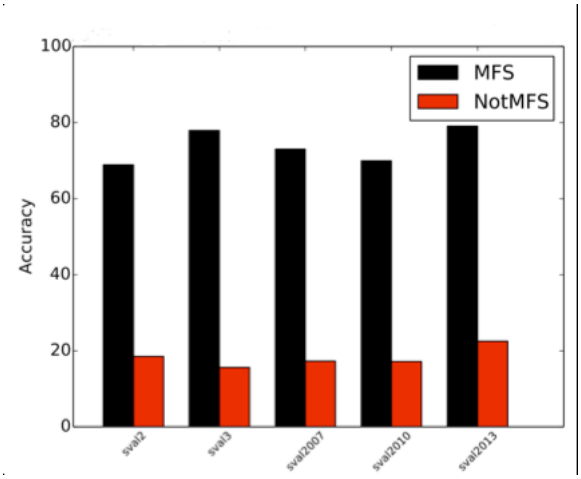

Fig. 2. Performance on most frequent and not most frequent sense cases

These issues are directly derived from the way that Natural Language Processing has been considered and the way in which NLP applications have been developed. These applications are framed mostly within computer science frameworks, in which it is relatively easy to define a specific task and an optimal expected output, but this is not so trivial in NLP. We propose to see Natural Language Processing as a big puzzle. The different tasks are small pieces that must fit perfectly in order to build an overall puzzle that represents the interpretation of a document or a text. Following the puzzle analogy, the pieces can not be considered in isolation. Moreover, sometimes external information is required to complete the puzzle, as for example knowing what is depicted in the puzzle to get clues about how to put the pieces together. Figure 3 shows the idea of the puzzle, where every NLP task is a small portion of the puzzle where all the

pieces must fit, but also the pieces of one task must fit with the rest of puzzle (rest of NLP tasks).

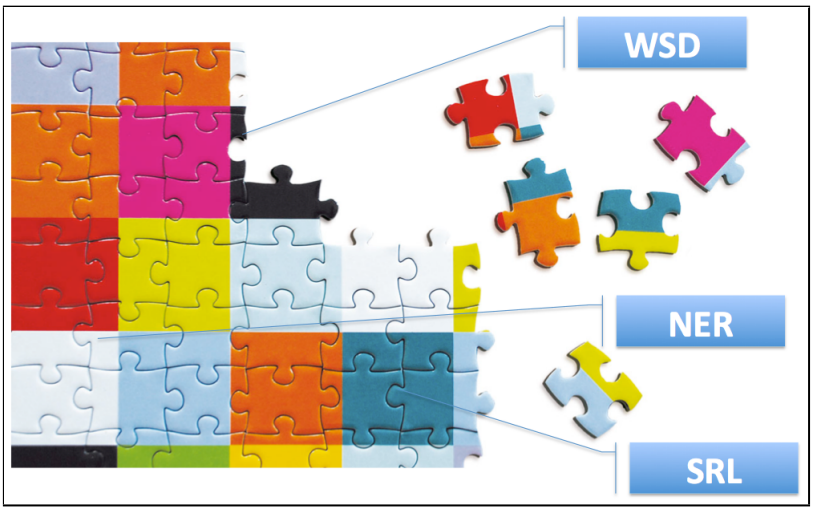

Fig. 3. The idea of the NLP tasks as a puzzle

Hence, the scope of the work is bringing together approaches that consider in different ways the hypothesis presented previously. For instance, approaches trying to solve several NLP task at the same time and mutually using the information among the specific subtasks to reach a good overall solution. Other interesting research would be using external knowledge resources (such as DBpedia, Wikipedia or the Web), in order to extract background and real–world information that could be used to understand texts and solve NLP problems.

This workshop has not been organized previously, but we think it deals with very relevant topics, which are being currently faced in a large range of NLP fields. It targets anybody working on Computational Linguistics and Natural Language applications and concerned with the ideas and approaches presented here. Some topics of interest could include among others:

- Natural Language applications
- Ambiguity resolution
- Global optimization
- Context modeling
- Use of external knowledge resources
- Word Sense Disambiguation
- Named Entity Recognition and Linking
- Named Entity Linking
- Event Extraction
- Event Coreference

Papers submitted to this workshop should address some of these points:

- Dealing with more than one NLP task
- Using background information, external sources and Linked Data
- Combining different external resources
- Modeling the context considering scopes larger than the sentence
- Processing multiple documents and linking information across them
- Influence of the domain and building domain specific resources to help NLP applications

References

1. Ruben Izquierdo, Marten Postma and Piek Vossen. *Error analysis of Word Sense Disambiguation*, In proceedings of CLIN2015: Computational Linguistics in The Netherlands, Antwerp, Belgium. February 2015.