PhD Thesis Proposal Review

Proposal title: Hybrid Machine Translation
Author: Mgr. Amir Kamran
Opponent: doc. Ing. Zdeněk Žabokrtský, Ph.D.

Thesis description

The aim of the proposal under review is to combine so called rule-based and statistical Machine Translation techniques, in order to achieve better MT quality.

The proposal consists of three sections. After a short introductory section, which gives motivation for hybrid MT, a section which summarizes related work follows (6 pages), and a section called “On going research and work plan” (5 pages).

Comments and evaluation

The general motivation behind the proposal is clear: it is well known that different MT approaches lead to different error distributions, so there is a chance for getting better results by combining them.

I can have no serious objections against the formal quality of the proposal. The proposal is written in good English (even if one can find several typos), is logically structured and is followed by the list of references.

As for the literature review, I think it is sufficient, given the limited size of the proposal. Personally, I would not recommend to follow the classical dichotomy *rule-based* vs. *statistical* any more, because I find it outdated in the contemporary MT world, if not even misleading. Strictly non-statistical MT systems are rather exotic nowadays. I understand that certain simplification is inevitable when describing such a broad field, but I believe that a more tangible categorization could be drawn, e.g. according to what data structure is used as the latent sentence representation in a given system (flat chunking into so called phrases, several sorts of trees, etc.).

I have two concerns about the last section, which I expected to describe author's own contributions, ideas and plans.

First, the thesis goals remain very vague even after the third section (let me cite: “We plan to extend the set of rules with a deeper understanding of the source and target languages.” – this sentence could embrace virtually anything). Moreover, I do not see any novel spots in the description of the goals. Mostly it is just mentioned that this or that technique (published about 5-10 years ago) will be used for this or that language pair. I am far from saying that general strategies such as using factored translation or introducing an artificial middle language have been already exhausted; definitely they haven't. But at this stage of the study, the student should already be able to present some thoughts of his own, and in a more concrete and elaborated way.

Second, there is almost no description of the student's own work (I mean what has been already done). Again, at this stage of study, he should be able to present at least some preliminary outputs (be they of empirical or theoretical nature), not just references to literature and future plans. I tried to search for cues in the list of references, but there is just one published paper co-authored by Amir Kamran, and there is a nonsensical reference to his non-existing PhD thesis (!).

Conclusion

Amir Kamran shows that he possesses a good orientation in the field of Machine Translation. However, in my opinion he should start elaborating his research goals in much more depth soon. Above all, he should try to identify their novelty potential in an explicit way.

In Prague, January 19, 2013

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University in Prague