Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis proposal

General Information

Master's Thesis Title:	**Subtitles translation for the English-Catalan language pair**
Orientation:	☐ professional ☒ research
M.Sc. Th. Advisor's Dept. & University:	LSI, UPC
M.Sc. Th. Advisor:	Lluís Màrquez, Cristina España
M.Sc. Th. Advisor e-mail:	lluism@lsi.upc.edu, cristinae@lsi.upc.edu
Observations:	
Student's Name: (if already known)	

M.Sc. Thesis Description

Main issues / Brief Description:

The goal of this thesis is to develop a subtitles translator based on open source translation software. The system will be applied to the English-Catalan language pair and made available for public use through a web interface.

Detailed Description:

Subtitle translation is a research field with gaining importance in countries where there is no tradition in dubbing TV programs or films [1,2]. From the point of view of the translation system, it is interesting because sentences are short and statistical engines can give good account of them. On the other hand, systems either need a large coverage to deal with every topic and different language styles, or have to be adapted to a restricted domain.

The work plan for this thesis involves five main steps:
- Become familiar with SMT techniques [3,4] and standard software, mainly Moses [5].
- Collect and prepare parallel corpora to train a statistical translator.
- Adapt the general translator to the subtitle task.
- Evaluate the performance of the resulting systems in different scenarios.
- Implement a web interface to make the system publicly available.

Research aspects of this work plan may include, but are not limited to:
- Sentence alignment of comparable corpora [6].
- Translation into a low-resource language. Direct translation vs. translation through a pivot language.
- Domain adaptation.

References:

[1] Volk, M (2008). The automatic translation of film subtitles: a machine translation success story? In: Nivre, J; Dahllöf, M; Megyesi, B. Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein. Uppsala, Sweden, 202-214.

[2] M. Volk, R. Sennrich, C. Hardmeier and F. Tidström (2010). Machine Translation of TV Subtitles for Large Scale Production. Proceedings of the Second Joint EM-CNGL Workshop "Bringing MT to the User: Research on integrating MT in the Translation Industry", 53-62.

[3] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2), 263-311.

[4] P. Koehn, F.J. Och, and D. Marcu (2003). Statistical phrase based translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL).

[5] P. Koehn et al. (2007), Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[6] Tiedemann, J. Improved Sentence Alignment for Movie Subtitles. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07), 2007.

Other comments:

Barcelona, October 22nd 2010