Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis proposal

General Information

Master's Thesis Title:	**High quality hybrid machine translation for specific domains**
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:	Research work within a project framework
Student's Name: (if already known)	

M.Sc. Thesis Description

Main issues / Brief Description:

The purpose of this work is to build a hybrid Machine Translation system joining Statistical Machine Translation techniques with Rule-based models. The resulting systems will be adapted to restricted domains such as patents and evaluated both in open and specific domains.

Detailed Description:

Statistical Machine Translation (SMT) [1,2] is a common paradigm for MT which offers robustness and flexibility, especially when one has a large amount of parallel texts available and adequate fragment translations can be obtained statistically. On the other hand, Rule-based Machine Translation (RBMT) [3] relies on linguistic rules and dictionaries to translate a sentence. For certain language pairs and constrained domains this approach can provide high quality translation.

The final goal of this thesis is to build a system with the best of each approach in an incremental way. The work involves:
- Become familiar with both SMT and RBMT techniques and the standard software, mainly Moses and GF [4,5].
- Build combination baselines from the available raw systems.
- Study different hybridisation strategies both/either led by the SMT system and/or the RBMT system.
- Evaluation of the resulting systems in open and restricted domains.

This is a pure research thesis, so the work plan is neither fix nor untouchable. It is however related to the work being done within the European project MOLTO [6], so there are some open research lines. In particular, for the final hybrid system we propose four possible approaches to explore:
- A straightforward approach: Force fixed GF partial translations within a SMT system.
More elaborated ways to integrate both strategies can be divided according to the main translation engine:
- Led by SMT: GF partial output, as phrase pairs, is integrated as a discriminative probability feature model in a phrase-based SMT system.
- Led by SMT: GF partial output, as tree fragment pairs, is integrated as a discriminative probability model in a syntax-based SMT system.
- Led by GF: Complement with SMT options the GF translation structure and perform statistical search to find the final translation.

References:

[1] 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.

[2] 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).

[3] W. J. Hutchins, and H.L. Somers. An introduction to Machine Translation. Academic Press, (1992)

[4] 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.

[5] A. Ranta. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. The Journal of Functional Programming 14(2), 145–189.

[6] http://www.molto-project.eu/

Other comments:

Barcelona, October 22nd 2010