Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis proposal

General Information

Master's Thesis Title:	**Syntactic and Semantic Features for Discriminative Phrase Selection in Statistical Machine Translation**
Orientation:	□ professional ☒ research
M.Sc. Th. Advisor's Dept. & University:	LSI,UPC
M.Sc. Th. Advisor:	Lluís Màrquez and Jesús Giménez
M.Sc. Th. Advisor e-mail:	lluism@lsi.upc.edu, jgimenez@lsi.upc.edu
Observations:	The master thesis is involved in a project
Student's Name: (if already known)	

M.Sc. Thesis Description

Main issues / Brief Description:

The goal of the project is to enhance an existing statistical Machine Translation (MT) system through the incorporation of syntactic and semantic features in the construction of translation models.

Detailed Description:

In standard phrase-based statistical MT systems [1], translation models are built on the basis of relative frequency counts, i.e., Maximum Likelihood Estimates (MLE). Thus, all the occurrences of the same source phrase are assigned, no matter what the context is, the same set of translation probabilities. For that reason, recently, there is a growing interest in the application of discriminative learning [2]. Discriminative translation models are able to take into account a wider feature context. Lexical selection is addressed as a classification task. For each possible source word (or phrase) according to a given bilingual lexical inventory (e.g., the translation model), a distinct classifier is trained to predict lexical correspondences based on local context. Thus, during decoding, for every distinct instance of every source phrase a distinct context-aware translation probability distribution is potentially available.

The aim of the project is to design specific syntactic and semantic features in an existing software for the construction of discriminative translation models. These features will be based on automatically obtained syntactic dependency trees and semantic role structures.

The thesis work will involve:

- Study of statistical MT [1]
- Study of the application of discriminative learning to translation modeling [2]
- Study of the MOSES MT system (http://www.statmt.org/moses/)
- Study of the MLT software for the construction of discriminative translation models.
- Design and implementation of syntactic and semantic features inside MLT.
- Construction and evaluation of syntactico-semantic discriminative translation models.
- (Possibly) Participation in the International Workshop in Machine Translation.

Other comments:

- Previous knowledge of the Perl programming language is desirable (not required)
- Knowledge of Machine Learning techniques is desirable (not required)

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2003.

[2] Jesús Giménez. and Lluís Màrquez:. Jesús Giménez and Lluís Màrquez. Discriminative Phrase Selection for Statistical Machine Translation. In Learning Machine Translation. NIPS Workshop Series. MIT Press. 2009.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Barcelona, October 18th 2010