Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis proposal

General Information

Master's Thesis Title:	**TITLE** Opinion Mining from a Large Corpora of Natural Language Reviews
Orientation:	☐ professional ☒ research
M.Sc. Th. Advisor's Dept. & University:	LSI, UPC
M.Sc. Th. Advisor:	Lluís Màrquez
M.Sc. Th. Advisor e-mail:	lluism@lsi.upc.edu
Observations:	
Student's Name:	Beltrán Borja Fiz Pontiveros

M.Sc. Thesis Description

Brief Description

The goal of this thesis is to develop a system for automatically processing a large database of textual hotel reviews in natural language to extract relevant opinions from users on a series of predefined features of quality (service, food, equipment, etc.). The information extracted has to be categorized according to polarity (positive/negative opinions), and arranged by hotel and topic to allow its usage from a hotel search application.

This work will imply the usage of state-of-the-art Natural Language Processing (NLP) technology for addressing several sub-problems, such as: parsing, named entity recognition, co-reference resolution, semantic equivalence, opinion detection, negation, uncertainty and speculative language, scope, etc.

Detailed Description

The framework of this master thesis is the QUESTIA project (TSI-020100-2010-210) involving ExperienceON and the Technical University of Catalonia (UPC), with the collaboration of the University of Oxford. This is a Spanish applied research project within the Avanza program, which aims at developing a search engine able to answer complex queries expressed in natural language. Key components of the system are: 1) a semantic parser able to process arbitrarily complex input queries in natural language, 2) a graph-based semantic representation language with concepts from a world ontology, and 3) a reasoning system wich is able to convert semantic representations into DB queries.
ExperienceON will develop an implementation of these ideas and deploy a hotel-search engine for commercial use. UPC will provide part of the Natural Language Processing Technology necessary to meet the objectives.

Within this framework, the master thesis goal is to develop a complementary module for the final application. In particular, the main goal is to develop a system for automatically processing a large database of textual hotel reviews in natural language to extract relevant opinions from users on a series of predefined features of quality (service, food, equipment, etc.). The information extracted has to be categorized according to polarity (positive/negative opinions) and arranged so that the final search application can use it to display complementary information of each hotel relevant to the user queries.

To have an idea of the usage of opinion mining over user reviews on hotels, one can consult the following website http://www.trustyou.com/

This master thesis work will imply the usage of state-of-the-art Natural Language Processing (NLP) technology for addressing several sub-problems, such as: parsing, named entity recognition, co-reference resolution, semantic equivalence, opinion detection, negation, uncertainty and speculative language, scope, etc. See the following references [1-8]

Main steps:
1) Familiarize with the existing bibliography
2) Initial study of the corpus of reviews
3) Usage of existing NLP analyzers to automatically annotate the reviews
4) Work on the opinion-mining specific module
5) Evaluation of results
6) Work on a initial presentation layer for the application
7) Documentation of the master thesis

References:

1. Bing Liu. "Opinion Mining." Invited contribution to *Encyclopedia of Database Systems*, 2008.
2. Bing Liu Sentiment Analysis and Subjectivity. Chapter in the *Handbook of Natural Language Processing*, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.
3. Bing Liu's website with general information on Opinion Mining http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
4. V. Stoyanov and C. Cardie, "Partially supervised coreference resolution for opinión summarization through structured rule learning," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 336–344, Sydney, Australia: July 2006.
5. Richárd Farkas; Veronika Vincze; György Móra; János Csirik; György Szarvas. *The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text.* Proceedings of CoNLL-2010. See the sared task website http://www.inf.u-szeged.hu/rgai/conll2010st/
6. Proceedings of the workshop "Negation and Speculation in Natural Language Processing", July 10, 2010, Uppsala, Sweden. http://aclweb.org/anthology-new/signll.html#2010_2
7. Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL-10), 2010.
8. (ongoing) Special Issue of the Computational Linguistics Journal on Modality and Negation. Roser Morante and Caroline Sporleder editors. http://cljournal.org/specials/modality-and-polarity.html

Other comments:

Barcelona, January 29th 2011