Introduction to the EXCITEMENT project: towards an open platform for EXploring Customer Interactions through Textual entailMENT

FP7 EC project, Language Technology Unit, 287923 – STREP
http://www.excitement-project.eu/
Started: January 2012

Moshe Wasserblat (coordinator)¹, Ezra Daya¹, Eyal Hurvitz¹, Maya Gorodetsky¹, Dmitri Volsky¹, Ido Dagan², Meni Adler², Asher Steren², Sebastian Pado³, Tae-Gil NOH³, Britta Zeller³, Günter Neumann⁴, Kathrin Eichler⁴, Rui Wang⁴, Gabriele Fidanza⁵, Giorgio Gianforme⁵, Matthias Meisdrock⁶, Magnini Bernadom⁷, Bentivogli Luisa⁷, Zanoli Roberto⁷, Lavelli Alberto⁷

¹ NICE Systems, Israel; ² Bar-Ilan University, Israel; ³ Heidelberg University, Germany; ⁴ DFKI, Germany; ⁵ ALMAWAVE, Italy; ⁶ OMQ, Germany; ⁷ FBK, Italy

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n°287923

I. ABSTRACT

Identifying semantic inferences between text units is a major underlying language processing task, needed in practically all text understanding applications. While such inferences are broadly needed, there are currently no generic semantic “engines” or platforms for broad textual inference. The primary scientific motivation for the EXCITEMENT project is to change this ineffective state of affairs and to offer an encompassing open source platform for textual inference. On the industrial side, EXCITEMENT is focused on the text analytics and speech analytics markets and follows the increasing demand for automatically analyzing customer interactions, which today cross multiple channels including speech, email, chat and social media.

II. INTRODUCTION

Performing inferences over texts is a major underlying task for exploiting natural language information, needed in practically all text understanding applications. For example, to identify the correct answer for the question “Who acquired Overture?” from the text snippet “Yahoo’s buyout of Overture was approved…”, a question answering (QA) system needs to recognize that the statement “Yahoo acquired Overture” can be inferred from the given text. Similarly, such inferences are needed in an information extraction (IE) task targeting the extraction of company acquisitions. In a different setting, multi-document text summarization systems need to recognize that the meaning of a sentence taken from one document can be inferred from the meaning of a differently-stated sentence in another document, in order to avoid including both sentences in the generated summary. Related settings include machine translation (MT) evaluation, where the task is to recognize that the meaning of an automatically-generated translation is semantically equivalent to a manually-generated reference translation, and automatic scoring of students’ free-text answers to open questions, where the intelligent-tutoring system needs to recognize that the student’s answer implies
the reference answer (i.e. the latter can be inferred from the former).

There are two interleaved high-level goals for this project, which would yield two corresponding outcomes. The first is to set up, for the first time, a generic architecture and a comprehensive implementation for a multilingual textual inference platform and to make it available to the scientific and technological communities. To a large extent, the idea is to follow the successful experience of the Moses open source environment for machine translation, which has been making a substantial impact on research in that field. This will enable developers of many text-processing applications to leverage the platform and boost their semantic inference capabilities. It will also provide developers of inference technology an effective environment for implementing and evaluating their components, and an easy entry-point for research in this field.

The second goal of the project is to develop a new generation of inference-based industrial text and speech exploration applications for customer interactions, which will enable businesses to better analyze and make sense of their diverse and often unpredicted client content. These goals will be achieved for three languages – English, German and Italian, and for three customer interaction channels – speech (transcriptions), email and social media.

III. CONTRIBUTIONS

This section provides details on the contributions that EXCITEMENT will make. We structure this section around two sets of contributions.

Scientific contribution concerns with advances on the scientific side of textual entailment, and can be summarized as leading to a generic, multilingual platform for component-based textual entailment research.

Industry contribution targets the applied side of textual entailment, and translates the scientific advances into robust, efficient textual entailment engines for practical business analytics and text exploration on an industrial scale.

Scientific contribution: A Multilingual Open Platform for Component-Based Textual Entailment

This set of contribution addresses the shortcomings of the current state of the art in textual entailment, in particular the fragmentation across approaches and languages. We make three concrete contributions: A component-based architecture which subsumes all existing work and lends itself to multilingual detection of entailment (A1), algorithmic advances that will lead to better knowledge acquisitions methods and language-independent inference algorithms (A2), and an implementation of the component-based architecture that will be made freely available and can serve as a “reference” textual entailment system, including knowledge resources for three languages (English, German, Italian) (A3).

Contribution A1: A Generic Multilingual Architecture for Component-Based Textual Entailment

The goal is to develop a modular API for which has language independent components, algorithms and resources. Figure 1, illustrates the EXCITEMENT architecture: the left-hand side (solid lines) shows the data flow at application time (for entailment detection), while the right-hand side shows the data flow at system development time (for knowledge acquisition). As the figure shows, we will separate entailment detection into independent sets of components: (a), linguistic analysis components which provide the analysis of text and hypothesis; (b), individual inference components which cover different sub-phenomena affecting entailment or provide entailment knowledge; (c), the inference algorithm proper that interacts with the inference components and computes the final entailment decision utilizing their outputs.
Contribution A2: A Algorithmic Progress in Multilingual Entailment Technology

The goal is to develop novel methods to improve knowledge acquisition within and across languages. In addition develop novel inference algorithms that improve robustness while maintaining precision.

Contribution A3: An Open-Source Multilingual Reference Implementation of Component-Based Textual Entailment

Develop an Open-source reference implementation of a textual entailment platform for practical exploitation of Goal A1 (Similar to MOSES in MT community) and encourage of an open source community around the platform.

The open source platform will be a de-facto standard which (a) can be used to kick off textual entailment projects quickly, (b), ensures interoperability of resources between different research groups, (c), allows for transparent and meaningful evaluation of different entailment setups in practice, and (d), can be extended by other groups with arbitrary new functionality. We are also strongly committed to establishing the viability of the platform beyond the lifetime of the EXCITEMENT project.

Industry contribution: A Multilingual of Textual Entailment for Text exploration

Interaction Analytics is a fast growing area of research with tremendous commercial potential. The processing and analysis of customer interactions in the business environment, using multilingual text-entailment (TE) technology, will enable automatic extraction of root-causes of business issues and new exploration. The TE infrastructure will ease the implementation of Speech and Text Analytics applications in the organization.

Our industry partners will provide data in three languages from three domains of economic interest, namely mobile service providers (English), telecommunications (Italian), and software (German). Furthermore, these real-world data stem from three different channels: speech transcriptions, email, and social media forums.

IV. IMPACT

The expected impact of EXCITEMENT is driven by the prospect to lay new grounds for powerful textual inference technology. On the scientific side, the realization of the generic textual entailment paradigm within an encompassing open platform would end the current state of affairs in applied semantics technology, which, for a long time was lacking a feasible unifying driving framework. On the industrial side, providing new inference capabilities will open new horizons for text analytics in general, and customer interaction analytics in particular, which will enable businesses to better harness the value of their customers’ inputs and thus increase their competitiveness.

Altogether, the expected outcome of EXCITEMENT ranges from new scientific insights all the way to novel practical technology in the hands of European developers and end-users. The consortium’s dedication to the open source platform, shared resources and transparent scientific dissemination will ensure that the results strengthen the R&D base in applied
semantics and text exploration: from university students to commercial developers, all types of both academic and industrial players will benefit from the proposed project.

V. SUMMARY

We believe that the open platform for component-based text entailment will be an important scientific contribution for Europe and will stimulate innovation and technology uptake. We expect that Textual Inferences components may become standard modules in concrete application scenarios, overcoming some of the drawbacks currently experienced by traditional approaches, among which the difficulty of maintenance of large lexicons and the non-reusability of the semantic components. Notably, Textual Inference solutions, when adopted in complex applications, will decrease the overall cost of system development, opening the market of semantic-based products to a larger spectrum of customers.

The market demand for solutions using these technologies will come from all public and private organizations that need to analyze and explore large amounts of unstructured information. The envisaged technology will enable the organization, their customers, and their employees to create value by better understanding the “voice of the customer” through multi-channel communications, such as telephone, e-mail, chats and social media.