An Emerging Transcontinental Collaborative Research and Education Agenda in Human Language Technologies

Gregory Ernest Monaco

Great Plains Network
Topeka, Kansas
and
University of Kansas
greg@greatplains.net

Abdelhadi Soudi

CLC and CS department
Ecole Nationale
de L'Industrie Minérale
Rabat, Morocco
asoudi@enim.ac.ma

Abstract

This paper describes an emerging interdisciplinary research and education collaboration, in the area of Human Language Technology (HLT), among multiple institutions in the Great Plains (US) and Morocco. In particular, this collaboration is focusing on the complex problem of meeting the language technology needs of societies with multiple dominant and indigenous languages, as in the United States and Morocco, by, among other things, developing language resources for the spoken languages (Berber, Moroccan Arabic and Native American languages) in the two countries. We present the goal of the project, the activities undertaken so far and the road ahead in this collaboration.

Introduction

By its very nature, Human Language Technology (HLT) is an interdisciplinary[1] enterprise involving, at a minimum, the fields of computer science, linguistics, psychology and (special) education, hence the necessity of collaboration between teams comprised of individuals with different specialties. When taking into consideration the actual domain of information under consideration (e.g., medicine, law, chemistry, physics) additional specialization may be required.

Consider the case of building an effective system for machine translation of text between languages that maximizes reader comprehension. We know that comprehension of information under ideal conditions (two native speakers of the same language, a relaxed setting) rarely, if ever, results in one to one mapping between the speaker's intent and the listener's understanding (Harris & Monaco, 1978). Two variables that affect comprehension of written information are text difficulty and skill of the reader (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998). It is not surprising that, for information that is translated by a computer (Machine Translation or MT) between languages or between sensory modalities, rates of comprehension drops precipitously due to characteristics of the author, accuracy of the computer program which does the translation, and, finally, characteristics of the reader (Murphy, 2000).

The factors involved in end-user comprehension of machine translated text are amenable to interdisciplinary analysis of the translation scheme, itself, plus an analysis of the host of human information-processing and text-representation variables ultimately involved in human comprehension.

It follows, then, that work on Human Language Technology applications, such as MT, prompts the need for the involvement of researchers with complimentary skills (computer science, linguistics, psychology, etc.).

It is in this context that a multidisciplinary and multi-institutional team from the Great Plains Network[2] together with a multidisciplinary and multi-institutional Moroccan

[1] We use the term *multidisciplinary* to describe the fact that the field of Human Language Technology covers many disciplines; we use the term *interdisciplinary* to describe the activity of individuals from diverse disciplinary backgrounds working together to identify and resolve common problems.

[2] Initially funded by the National Science Foundation to provide advanced networking services to member institutions in the seven US states of Arkansas, Kansas, Missouri, Nebraska, Oklahoma, North Dakota and South Dakota, the mission of the Great Plains Network (http://www.greatplains.net, http://research.greatplains.net) includes the establishment of large-scale joint projects, which are of interest to significant portions of the higher education membership in the region.

team have been intensively coordinating the establishment of a collaborative, interdisciplinary research and education agenda in HLT to work on HLT applications involving Arabic (standard and Moroccan), Berber, English and Native American languages and to initiate a joint Masters degree in HLT.

1. Goal of the Joint Research and Education Project

The goal of our joint project is to:

- Identify a set of common interdisciplinary research issues and a joint research plan in Human Language Technology, relevant both to spoken and written languages (Standard Arabic, Moroccan Arabic, Berber and Native American languages), and

- Initiate development of a joint program of study in HLT, which is interdisciplinary, takes advantage of developments in distance education to initially team scholars from the two countries to develop curricula and to teach together, and involves human resource sharing (students, faculty) between the Great Plains and Morocco.

2. Current Status of the Project

The current status of the project consists in the identification and prioritization of potential:

- Language resources (LRs) to be developed for the above cited languages. In this context, we plan to coordinate on this activity with :

(i) The Network of Euro-Mediterranean Language Resources (NEMLAR) project partners whose goal is to create a network of Euro-Mediterranean partners to specify and support the development of high priority LRs for Arabic and other local languages;

(ii) The Institute of Computational Linguistics (Pisa, Italy) and the “Ecole Nationale de l’Industrie Minerale” (Rabat, Morocco) who will be working on the development of LRs and tools that support MT between Arabic and Italian within the framework of a cooperation agreement between the National Center for Scientific and Technical Research (Morocco) and the National Center for Research (Italy);

(iii) The “Ecole Nationale de l’Industrie Minerale” (Rabat, Morocco) and a multi-institutional team who have been working on the development of Standard Arabic generation components for machine translation to Arabic. This project is supported by the National Center for Scientific and Technical Research (Morocco);

(iv) The “Institut d’Etudes et de Recherche sur l’Arabization” (Rabat, Morocco) which has a long-standing experience in Arabic Linguistics; and

(v) The Royal Institute of Tamaghizt (Rabat, Morocco) whose specific mission is the promotion of Berber.

We are identifying the LRs required for the two research project proposals cited below.

- Application areas (e.g., machine translation, information extraction, Cross-Language Information Retrieval). Two research project proposals are being examined:

(i) an English-Arabic MT system in the domain of finance and

(ii) Automatic Simplification of Arabic Text for Impaired Learners and Foreign Language Learners (Monaco & Smith, 1991, 1989; Soudi, 2004; Soudi & Eisele, 2004; Marsi et al., 2003; Soudi et al., 2003; Soudi et al. 2002b; Soudi et al. 2002a; Cavalli-Sforza et al., 2001).

In order to reinforce the research agenda, the GPN and Moroccan partners have been exploring possibilities for initiating a joint Master's degree program in HLT. The courses as well as most of the teaching faculty have been identified. Further details on this joint education program will be available at the Ecole Nationale de l’Industrie Minerale website soon.

While much of the initial work on the above initial activities has been accomplished, a meeting of collaborators to define scope of research, language resources to be developed, responsibilities and timeline is planned for the third quarter of 2004.

3. Collaborators

3.1. US Collaborators:

(i) The Great Plains Network (GPN): The GPN membership (21 universities) represents a diverse pool of talent in the areas of computer science, linguistics, psychology, cognitive science, education and special education, transfer of training, and computer application development. The GPN has also experience in building collaborative teams.

(ii) The American Distance Education Consortium: the involvement of the American Distance Education Consortium will also bring international experience in adult distance education and existing collaborations with Native American tribal colleges as well as expertise especially relevant in the area of development of a joint educational agenda.

(iii) Monaco & Associates Incorporated, a private US corporation that is investigating the joint development of products for commercial use. Monaco & Associates has extensive experience in project management for applied product development (e.g., Monaco & Smith, 1989, 1991; Monaco & Tomiser, 1992, 1995, 1996, 1999; Monaco & Wu, 1999, Monaco & Chang, in progress) as well as augmentative systems and rehabilitation.

3.2. Moroccan Collaborators:

(i) A multi-institutional and multidisciplinary team with expertise in the following areas:

- Computational morphology
- Speech synthesis
- Cross-language information retrieval
- Machine translation
- Text Categorization
- Linguistics
- Terminology

(ii) The Moroccan National Center for Scientific and Technical Research (CNRST): CNRST has a strong interest in the development of HLT applications that involve Standard Arabic and the local spoken languages. CNRST has recently funded two research projects proposed by the Ecole Nationale de l'Industrie Minerale, the coordinating partner in Morocco of this joint research and education agenda.

It is the long-term intention to extend the collaboration beyond the Great Plains and Morocco, to other national and international teams with complimentary skills and/or common interests.

Conclusion

In this paper, we have sketched an emerging interdisciplinary research and education collaboration, in the area of Human Language Technology (HLT), among multiple institutions in the Great Plains (US) and Morocco. We have outlined the goal of the project, the activities undertaken so far and the road ahead in this collaboration.

References

Cavalli-Sforza, V., Soudi, A. and Mitamura, T. (2000): Arabic Morphology Generation Using a Concatenative Strategy. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, Seattle, April 29-May 3, pp. 86-93.

Harris, R.J., & Monaco, G.E. (1978) The psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology*: *General*, *107*: 1-22.

Marsi E., Soudi, A., and Van den Bosch, A (2003) "Memory-based Arabic Morphological Analysis". *In Proceedings of the 14th Conference of Computational Linguistics in the Netherlands*, Avers, Belgique, December 2003.

Monaco, G.E., & Smith, J.P. (1989) *Basic Concepts for Teaching People with Developmental Disabilities and Mental Retardation.* Series of 10 Videotapes. Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Smith, J.P. (1991) *Challenging Issues in Developmental Disabilities.* Series of 8 Videotapes & 8 Program Guides. Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Tomiser, J.M. (1999) *EC3 Human Service Software.* Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Tomiser, J.M. (1996) *EC2 Service Coordination Software.* Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Tomiser, J.M. (1995) *EnCompass Service Coordination Software.* Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Tomiser, J.M. (1992) *MASSoftware*. Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Wu, L. (1999) *iEC Human Service Software for the Internet.* Topeka: Monaco & Associates Incorporated.

Monaco, G.E., & Chang, K. (*in progress*) *EConline*. Topeka: Monaco & Associates Incorporated.

Murphy, D. Keeping Translation Technology under Control. *Machine Translation Review*, No. 11, 7-10, 2000.

Soudi, A. (2004) "Challenges in the Generation of Arabic sentences from Interlingua representations", *To be published in Proceedings of conference "Traitement Automatique des Langues Naturelles"*, Fez, Maroc, April, 2004.

Soudi, A. and Eisele, A. (2004). « Generating an Arabic Full-form Lexicon for Bidirectional Morphology Lookup ». To be published in *Proceedings of Language Resources Evaluation Conference (LREC'2004)*, Lisbon, Portugal, May 2004.

Soudi, A., Cavalli-Sforza,V. (2002c): Arabic Morphology Generation: A two-step strategy. In *Proceedings of the Arabic and Information Technology International Conference organized by "Le Haut Conseil de la Langue Arabe"*, Algiers, Algeria, 28-29 December 2002.

Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2002b): A Prototype English-to-Arabic Interlingua-based Machine Translation System. *In Proceedings of The Processing of Arabic Workshop, Language Resources Evaluation Conference (LREC'2002)*, Las Palmas, Spain, June 2002.

Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2002a): Arabic Noun System Generation. In *Proceedings of The International Conference on the Processing of Arabic*, Lamanouba University, Tunisia, April 2002, pp. 69-87.

Soudi, A., Cavalli-Sforza,V. and Jamari, A. (2001): A Computational Lexeme-Based Treatment of Arabic Morphology. In *Proceedings of the Association for Computational Linguistics, Arabic Processing Workshop*, Toulouse, France, July 2001, pp.155-162.

Wolfe, M.B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., & Landauer, T.K. Learning from Text: Matching Readers and Texts by Latent

Semantic Analysis. *Discourse Processes*, *25*: 309-36, 1998.