Collaboration: interoperability between people in the creation of language resources for less-resourced languages

A Speech and Language Technologies for Minority Languages (SALTMIL) workshop

Tuesday, May 27, 2008

Briony Williams, Bangor University, Wales
Mikel L. Forcada, Universitat d'Alacant, Spain
Kepa Sarasola, Euskal Herriko Unibertsitatea / University of the Basque Country

ABSTRACTS

WORKSHOP PROGRAMME
May 27, 2008

14:30-14:35 Welcome

14:35-15:05 *Icelandic Language Technology Ten Years Later*
Eiríkur Rögnvaldsson

15:05-15:35 *Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources*
Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych

15:35-16:05 *Building resources for African languages*
Karel Pala, Sonja Bosch, and Christiane Fellbaum

16:05-16:45 COFFEE BREAK

16:45-17:15 *Extracting bilingual word pairs from Wikipedia*
Francis M. Tyers and Jacques A. Pienaar

17:15-17:45 *Building a Basque/Spanish bilingual database for speaker verification*
Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Juan J. Igarza and Inma Hernaez

17:45-18:15 *Language resources for Uralic minority languages*
Attila Novák

18:15-18:45 *Eslema. Towards a Corpus for Asturian*
Xulio Viejo, Roser Saurí and Angel Neira

18:45-19:00 Annual General Meeting of the SALTMIL SIG

Icelandic Language Technology Ten Years Later

Eiríkur Rögnvaldsson; Department of Icelandic, University of Iceland, Reykjavík, Iceland

We describe the establishment and development of Icelandic language technology since its very beginning ten years ago. The ground was laid with a report from a committee appointed by the Minister of Education, Science and Culture in 1998. In this report, which was delivered in the spring of 1999, the committee proposed several actions to establish Icelandic language technology. This paper reviews the concrete tasks that the committee listed as important and their current status. It is shown that even though we still have a long way to go to reach all the goals set in the report, good progress has been made in most of the tasks. Icelandic participation in Nordic cooperation on language technology has been vital in this respect. In the final part of the paper, we speculate on the cost of Icelandic language technology and the future prospects of a small language like Icelandic in the age of information technology.

Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources

Heather Simpson[1], Christopher Cieri[1], Kazuaki Maeda[1], Kathryn Baker[2], and Boyan Onyshkevych[2]; [1]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA, [2]U.S. Department of Defense

The REFLEX-LCTL (Research on English and Foreign Language Exploitation) program, sponsored by the United States government, was a medium-scale effort in simultaneous creation of basic language resources for several less commonly taught languages (LCTLs). To address some of the gaps in language technologies and resources, and to spur new research in this area, two REFLEX-LCTL sites constructed language packs for 19 LCTLs, and distributed them to research and development also funded by the program. This paper will focus on the work done at the Linguistic Data Consortium (LDC). LDC created language packs for 13 out of the 19 languages: Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, and Yoruba. Discussed are the goals and reasoning behind the language choice and language pack construction, and more in depth on the human resource and technology challenges in creating these language packs.

Building resources for African languages

Karel Pala[1], Sonja Bosch[2], Christiane Fellbaum[3]; Faculty of Informatics, [1]Masaryk University, Brno, Czech Republic, [2]Department of African Languages, University of South Africa, Pretoria, South Africa, [3]Department of Psychology, Princeton University, Princeton, NJ, USA

We report on work towards the creation of African Languages WordNet, comprised of interlinked semantic networks in several African languages that are known to have limited language resources. Adding these languages to the WordNet family will enable NLP applications for each of the languages in isolation. Moreover, linking the African Wordnets to one another and to the many global WordNets will make crosslinguistic information retrieval and question answering possible, and significantly aid machine translation. In this paper it is demonstrated how collaborative work between people, using existing tools, can contribute to the building of large text corpora and subsequently address the challenge of limited availability of language resources. The long term aim is the development of aligned WordNets for Bantu languages spoken in South Africa as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa.

Extracting bilingual word pairs from Wikipedia

Francis M. Tyers[1,2], Jacques A. Pienaar[3]: [1]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Alacant, Spain, [2]Prompsit Language Engineering S.L., Sant Vicent del Raspeig, Spain; [3]Centre for Text Technology, North-West University, Potchefstroom 2531, South Africa

A bilingual dictionary or word list is an important resource for many purposes, among them, machine translation. For many language pairs these are either non-existent, or very often unavailable owing to licensing restrictions. We describe a simple, fast and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the interwiki link system) and assess the performance of this method with respect to four language pairs. Precision was found to be in the 69-92% region, but open to improvement.

Building a Basque/Spanish bilingual database for speaker verification

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Juan J. Igarza and Inma Hernaez, AhoLab Signal Processing Group, Department of Electronics and Telecommunications, University of the Basque Country, Bilbao, Spain.

Research groups aiming to record new speech databases for minority languages have to face a series of difficulties, such as the lack of previous resources, the scarcity of fluent speakers and the shortage of funding for the project. Sometimes it is possible to take advantage of recording campaigns for other projects, and extend them in order to make some recordings in that language for every contributor that is able to speak it. In this way, new databases can be recorded with little extra effort, as the campaign is already prepared and funded. Using this technique, a new Basque/Spanish bilingual database has been created. Thanks to this database, some research is being undertaken on Basque/Spanish bilingual speaker verification systems. We present a description of the resulting database and the difficulties encountered during its acquisition.

Language resources for Uralic minority languages

Attila Novák; MorphoLogic, Budapest, Hungary

Most members of the Uralic language family are small minority languages spoken on the territory of the Russian Federation, which all are endangered. In past and ongoing projects, computational morphologies and annotated corpora have been and are being created for several of these Uralic minority languages: Udmurt, Komi-Zyrian, Eastern Mari, Northern Mansi, and the Kazym and Synya dialects of Khanty, Tundra Nenets and Nganasan. This article presents the morphological analyzers and other annotation tools and the resources developed and used during the projects.

Eslema. Towards a Corpus for Asturian

Xulio Viejo[1], Roser Saurí[2], Ángel Neira[3]; [1] Departamento de Filología Española, Universidad de Oviedo, Oviedo, Spain, [2] Computer Science Department, Brandeis University, Waltham, Massachusetts, USA. [3] Computer Science Department, Universidad de Oviedo, Oviedo, Spain

We present Eslema, the first project devoted to building a corpus for Asturian, which is carried out at Oviedo University. Eslema receives minor funding from the Spanish government, which is fundamental for basic issues such as equipment acquisition. However, it is insufficient for hiring researchers for a reasonable period of time. The scarcity of funding prompted us to look for much needed resources in entities with no institutional relation to the project, such as publishing companies and radio stations. In addition, we have started collaborations with external research groups. We are for example initiating a project devoted to developing a wiki-based platform, to be used by the community of Asturian speakers, for loading and annotating texts in Eslema. That will benefit both our project, allowing to enlarge the corpus at a minimum cost, and the Asturian community, causing a stronger presence of Asturian in information technologies and, as a consequence, boosting the confidence of speakers in their language, which will hopefully contribute to slow down the serious process of substitution it is currently undergoing.