The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

January - December 2012

Vol.16 n.1 & 4

LREC 2012 Special Issue

8th International Conference on Language Resources and Evaluation

ISSN: 1026-8200

ELRA/ELDA
Secretary General: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.org
Web sites:
http://www.elra.info or
http://www.elda.org

Contents

Dear Colleagues,

Like every two years, the special issue of the ELRA newsletter is devoted to the Language Resources and Evaluation Conference (LREC). The Eighth edition of the Language Resources and Evaluation Conference took place last May in Istanbul (Turkey) under the Patronage of Ms Neelie Kroes, Vice-President of the European Commission, Digital Agenda Commissioner and Mr Nihat Ergün, Minister of Science, Industry and Technology of the Republic of Turkey.

This edition has been again very popular: 1225 participants coming from 65 countries registered to the main conference, workshops and tutorials. This time, Germany brought the highest number of participants, but the participation from America and Asia remained strong.

For LREC 2012, a new feature has been introduced:

The *Language Library* is an initiative which is conceived as a facility for gathering and making available the linguistic knowledge the field is able to produce, through simple functionalities, putting in place new ways of collaboration within the Language Resource. The Language Library is the first step towards a community-built space where the entire LRT community shares data about language resources and annotated/encoded language data.

And the features introduced in 2010 have been extended:

• The *LRE-Map*, a mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources during the submission process. More than 1200 language resource forms in more than 200 languages were collected at this LREC.

• The *EU Village*, an initiative supported by the European Commission to encourage EC-sponsored projects to gain visibility by showing their objectives, progress and activities, either through demos, or through brochures or posters for projects still at the early stages.

• The *Special sessions*, which are oral sessions on "hot topics", with a lower number of papers presented to leave room for more exchange and discussions between the authors, chairpersons and audience. In 2012, a long special session has been dedicated to EC Projects with the objective to provide a broad overview of the 60 LT projects launched by the European Commission during the past few years.

A CoCoFLaRE workshop on Reinforcing International Collaboration in LRE was held on May 26th 2012. This meeting was organised jointly by COCOSDA, the International Committee for Co-ordination and Standardisation of Speech Databases, and FLaReNet aimed at sharing views on ways to reinforce international collaboration in the field of Language Resources and Evaluation at a time when data sharing is coming to the forefront of science, in many different areas such as biology, astrophysics, humanities, etc. The discussion was structured around the main topics "Spoken language resources and evaluation" and "Enlarging the scope to speech & language in general".

Six years ago, the ELRA Board created the Zampolli Prize, a prize for "Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation", to honour the memory of its co-founder and first president, Antonio Zampolli. In 2012, the Antonio Zampolli Prize was awarded to **Charles Fillmore** and **Collin Baker**, from the International Computer Science Institute (ICSI), University of California Berkeley (USA), and **Oriental Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (Oriental COCOSDA)**.

The next edition of LREC will be held in Reykjavik (Iceland) on May 26 - June 1, 2014.

Now concerning the content of this ELRA newsletter dedicated to LREC 2012, a few sessions' summaries are proposed along with the Opening Ceremony speeches. A short report on participants' feedback is also available.

Last but not least, the new resources added to the ELRA catalogue are listed at the end of this newsletter.

Nicoletta Calzolari, President

Khalid Choukri, Secretary General

INTRODUCTION
by Nicoletta Calzolari, LREC 2012 Conference Chair

I wish first to express to Ms Neelie Kroes, Vice-President of the European Commission, Digital agenda Commissioner, the gratitude of the Program Committee and of all LREC participants for her Distinguished Patronage of LREC 2012.

Even if every time I feel we have reached the top, this 8th LREC is continuing the tradition of breaking previous records: this edition we received 1013 submissions and have accepted 697 papers, after reviewing by the impressive number of 715 colleagues. We have accepted 30 Workshops and 10 Tutorials. We can't deny that the field of Language Resources and Evaluation is flourishing more than ever! So far more than 1100 people have already registered. From all these signals we see that LREC continues to be - as many say - "the conference where you have to be and where you meet everyone".

The high acceptance rate is an important characteristic that is at the essence of LREC. We consider of utmost importance to provide a global view of the current trends, in all their dimensions and as reflected in many languages. Multilingualism is a core feature of the field and it is important to show how the field is evolving not only with respect to new methodologies and algorithms but also with respect to more advanced treatment of more and more languages. This is also a strategic choice underlying the importance of safeguarding the heritage of world's linguistic diversity.

In the preparation of the program, while trying to arrange all the pieces of the big puzzle, it is a pleasure touching the hot topics of these days. It is always extremely interesting to spot the major changes with respect to previous editions and monitor the evolution of the field.

Major trends, i.e. not the most crowded topics but those increasing with respect to last LREC:

- "Data", as normal in our conference, but I would say even more than before: data/corpora (in all the modalities and for many purposes: annotation, extraction, classification, translation, and so on) receive even more attention than last time.

Nicoletta Calzolari

- Machine Translation and multilingualism is a major topic, on which a lot of work is carried out.
- Infrastructural initiatives, strategies, national and international projects are a big issue, as usual inside the LREC community.
- Tools and systems for text analysis at many levels are presented in many papers.
- Temporal and spatial information is also relatively increasing.

New entries, i.e. topics emerging this year:

- New media (twitter, chats and the like)
- Crowdsourcing
- Child language corpora

More or less stable, at the same level as last year are topics such as:

- Semantics and Knowledge, in all their variations: from annotation of anaphoric information, to ontologies and lexicons, disambiguation, named entities recognition, information extraction, to mention just a few.
- Subjectivity, declined in various nuances: emotions, opinions, and sentiments.
- Dialogue and discourse, with contributions from both the Speech and Text communities.
- Speech and Multimodal resources, tools, systems.
- And finally, evaluation and validation methodologies, as an important part of quite many papers.

A slightly declining tendency seems to be associated with:

- Lexical resources
- Grammar, syntax, parsing

Are these solved issues?

As usual, a distinctive feature of LREC is the emphasis given to infrastructural and strategic initiatives. I consider this a very important characteristic of the Language Resource field, one that ELRA has always supported, and one that deserves some reflection. This is probably due to the recognised fact of the necessity to work on massive amounts of data, moreover multiplied by the many languages, and to the infrastructural nature of language resources. The fact that ours is a data-intensive discipline requires building on each other results if we want to achieve something serious, and requires joining efforts. This recognition has led in the last

years to the creation of important infrastructures, such as in Europe META-SHARE and CLARIN-ERIC.

We must now seriously think at how to enter with force into the area of "big data" - and also "open data" - the next frontier for competition and innovation, where we have to cope with very strong and organised communities. But we must also learn from them, in particular how to enable working together in huge experiments with thousands of colleagues cooperating on common objectives. This is, I believe, our next step if we want to address the challenges of big data and achieve the status of a mature science. And this requires networking, collaboration and sharing, of ideas and data. I hope LREC helps going in this direction.

LREC Innovations

We continue also the tradition of introducing some innovation at LREC.

- The ***LRE Map***, which started only two years ago, is already established, consulted every day and is used in other major conferences. At this LREC we collected descriptions for more than 1200 resources in more than 200 languages! More details on the Map will be found at the ELRA booth where it is presented.

- The novelty of this year is the ***Language Library***. It is an experiment to see if we can set up a platform for collaborative work on the processing (annotation, translation, ...) of language resources. The Library will be presented in the ELRA booth as well. Both the LRE Map and the Library have been conceived as services for the community and are being built by the community itself.

- Also this time we have introduced in the program two ***"Special sessions"*** on emerging new topics - New Media and Crowdsourcing - with a slot dedicated to general discussion.

- We repeat this year the experience of the ***EC Village***, so successful last time. We have even more booths, representing a very large number of EC projects.

- A new insertion of this LREC is the ***EC Track*** on the second day of the conference, organised by Roberto Cencioni, where some EC projects - and related trends - in the field of Language Resources and Evaluation are presented and discussed. This, together with the EC Village, offers the possibility of getting a comprehensive picture of the EC initiatives in the field.

Acknowledgments

And finally, I wish to express my appreciation to all those who made this LREC possible and hopefully successful.

I first thank the Program Committee members, not only for their dedication in the huge task of selecting the papers, but also for the constant involvement in the various aspects around LREC. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program. And obviously to Khalid Choukri, who is in charge of so many organisational aspects around LREC.

I thank ELRA, which is the promoter of LREC, its own conference.

Furthermore, on behalf of the Program Committee, I thank our impressively large Scientific Committee. They did a wonderful job.

A particular thanks goes to the Local Committee, and especially to Mehmed Özkan (its chair): they have worked hard for many months to find the best solutions to local issues.

I express my gratitude to the Sponsors that have believed in the importance of our conference, and have helped with financial support. I am grateful to the authorities, and all associations, organisations, companies that have supported LREC in various ways, for their important cooperation.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events. A big thanks goes to all the authors, who provide the "substance" to LREC, and give us such a broad picture of the field.

I finally thank the two institutions that have dedicated such a great effort to this LREC, as to the previous ones, i.e. ELDA in Paris and my group at ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least, thanks are thus for: Hélène Mazo and Sara Goggi, two pillars of LREC without whose commitment for many months LREC would not happen, and the others who have helped and will help during the conference: Victoria Arranz, Cécile Barbier, Paola Baroni, Roberto Bartolini, Riccardo Del Gratta, Francesca Frontini, Olivier Hamon, Valérie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller. We have solved together the many big and small problems of such a large conference. They will also assist you during the conference.

Now LREC is yours. You - the participants - are the real protagonist of LREC, you will make this LREC great. So, at the very end, my greatest thanks go to you all. I may not be able to speak with each one of you during the Conference (although I'll try!). I hope that you learn something, that you perceive and touch the excitement, fervour and liveliness of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation, which you will show at the next LREC.

I particularly hope that funding agencies all over the world will be impressed by the quality and quantity of the initiatives in our sector that LREC displays, and by the fact that the field attracts practically all the best groups of R&D from all continents. The success of LREC for us actually means the success of the field of Language Resources and Evaluation.

The tradition of holding LREC in wonderful locations with a Mediterranean flavour continues, and Istanbul is a perfect LREC location! I am sure you will enjoy Istanbul during the LREC week. And I hope that Istanbul will enjoy the invasion of LRECers!

With all the Programme Committee, I welcome you at LREC 2012 in such a wonderful country as Turkey and wish you a fruitful Conference.

Enjoy LREC 2012 in Istanbul!

Nicoletta Calzolari Zamorani
Istituto di Linguistica Computazionale del CNR
Via Moruzzi 1
56124 Pisa, Italy
glottolo@ilc.cnr.it

LREC 2012 Opening Ceremony Speeches

Message from Stelios Piperidis, the ELRA President

Stelios Piperidis

Welcome to LREC2012! Welcome to Istanbul!

Let me first express, on behalf of the ELRA Board and Members, our profound gratitude to Mrs Neelie Kroes, Vice-President of the European Commission, for her Distinguished Patronage of LREC 2012 and for honoring us with her message.

The 8th edition of LREC takes place in a most interesting context for our field, in times when everything is changing at a dazzling speed. The common denominator of all changes being the quest of the best recipe for rapid development, competitiveness and innovation. Information technology at large has an undisputable contribution to this endeavour, as access to the whole spectrum from data, information, content, up to knowledge, and the ability to process it is one of the enabling factors leading to development and innovation. Taking into account the importance of language in this spectrum, language technologies have a decisive role to play. To be able to fulfill the role, understand the complex phenomena, tackle open problems, build useful applications and foster innovation, access to the necessary resources infrastructure is indispensable. It is, thus, not to wonder why during the last years issues relating to data access, open data, collaboration and exchange have absorbed a non-negligible portion of the energy of the scientific world and certainly of our field.

ELRA, established already in 1995, has played a pioneering role in this respect. At times when data-driven techniques had just started rising, and numerical and learning methods had not prevailed in the language technology field, ELRA was set up as an association that would take care of identifying, archiving, and distributing language data, as well as cater for language technology evaluation. These initial goals were soon extended into production and validation of language data, production of technology evaluation packages, as well as into offerings of specific services and organisation of specific events, LREC being the major such event since 1998.

ELRA has achieved to establish itself as a high-quality data centre, as a main player in the field of language resources and evaluation, enjoying a steady base of membership. Throughout its life, the Association has been constantly responding to new challenges and needs of the field, as well as adapting to new cultures and trends. Besides its ever evolving catalogue of language data with clear distribution rights and licences, ELRA has set up the Universal Catalogue, a bottom-up, community-built resource radar, a catalogue of resources, data and tools, irrespective of whether these can be acquired through ELDA, ELRA's Distribution Agency, or not. It supports the LRE Map, a concerted effort towards a new culture, a rich, community-built, information base, already embraced and re-used by many conferences, a unique tool for monitoring progress, identifying specific gaps, highlighting trends. It also supports the emerging Language Library, a platform for collaboratively built annotated resources at all levels, available as open resources to the whole community. A big "thank you" to all participants for contributing to this initiative.

In a similar vein, through its catalogues and other initiatives, ELRA reinforces its cooperation links with major data centres around the world, notably LDC and NICT, trying to unify all component catalogues into a Global Inventory of appropriately identified LRs.

To achieve its goals, ELRA has restructured its activities along three axes : a) *infrastructural*, taking care of identification, cataloguing, updating metadata-based LR description and documentation, and new simpler distribution mechanisms, b) *scientific*, taking care of LRs production & validation, and evaluation, and c) *promotion*, taking care of marketing and promotion activities like LREC and information aggregation services.

Along these axes, ELRA has currently set 5 main priorities:

- Opening up segments of its catalogue

Anticipating users' expectations, ELRA has decided to offer a large number of resources for free for the research community. Such an offer will consist of several sets of speech, text, multimodal databases that will be released for free regularly, as soon as legal aspects are cleared.

- Fostering the use of Public Sector Information

Being among the first to recognize the importance of public sector information, the Association joins forces with all stakeholders in the effort to reinforce and extend the free use of public sector information for research, technology and application development.

- Supporting META-SHARE, the new resource sharing and exchange infrastructure

ELRA, through its operational body ELDA, is a founding member of META-NET, and plays an important role in META-SHARE offering a range of services (among others supporting the legal helpdesk and the non-local repository of the infrastructure) and helping sustain it. Most of the resources already distributed by ELRA are also available on META-SHARE.

• Promoting collaborative and crowd-sourcing based methods of language resources building

During the last year, the Association has been investigating the setup of a dedicated platform for crowd-sourcing based language resource building. Your feedback through the questionnaire we have prepared is most valuable.

• Establishing the Language Resources and Evaluation Forum (LRE-F)

ELRA is establishing the Forum that is expected to help maintain and extend our vibrant community through sharing, exchange and collaboration assisting services.

Just a few words about the Antonio Zampolli Prize, the prize created by the ELRA Board in order to honour our founder and first president who did so much for the field of language resources. Citing the Prize articles: "The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation. In awarding the prize we are seeking to reward and encourage innovation and inventiveness in the development and use of language resources and evaluation of HLTs. The prize covers the field of Language Resources and Language Technology Evaluation in the areas of spoken language, written language and terminology". At the LREC 2012 conference, the Prize will be awarded for the fifth time. The ELRA Board has been very happy to receive very strong nominations made by outstanding people in the field, and we do recognize there are several persons who are eligible for this prestigious prize.

I would like to take the opportunity to thank all those who have worked so hard to make this conference a big success: the LREC Programme Committee, chaired by Nicoletta Calzolari, the Scientific Committee, the International Advisory Committee, the group in Pisa, Khalid Choukri and the ELDA staff in Paris, the two LREC "pillars" Helene Mazo and Sara Goggi, Mehmed Özkan and the Local Organising Committee. Each one of them in his/her own role has been taking care of the myriad of issues that pop up when undertaking the organisation of such a complex and demanding conference as LREC, a Herculean task. Particular thanks to all sponsors and supporters, all conference participants, workshop and tutorial organizers, project consortia participating in the EC Village; you have all helped outperform once again all previous editions in all the dimensions of our LREC conference.

Dear LREC Participants, The Conference is now in your hands. With your active participation in the oral sessions, your lively discussions with the presenters at the poster sessions, your visits to the EC Village and participation in the EC Track to discuss with the investigators of the research projects funded by the European Commission, LREC 2012 will be yet another success.

I welcome you all to LREC 2012 in magnificent Istanbul and wish you a fruitful conference.

Stelios Piperidis
ILSP / R.C. "Athena"
Epidavrou & Artemidos 6
Marousi 151 25 Athens, Greece
spip@ilsp.gr

Message from Khalid Choukri, ELRA Secretary General and ELDA Managing Director

Welcome to Istanbul and LREC 2012,

Mr. Nihat Ergün, Minister of Science, Industry and Technology of the Republic of Turkey (left) and Khalid Choukri (right)

Welcome to this LREC 2012, the 8th edition of one of the major events in language sciences and technologies and the most visible service of ELRA to the community.

I would like to extend our warm welcome to the 140 representatives of ELRA members, attending LREC2012.

On behalf of ELRA members and LREC participants, I would like express our gratitude to Ms Neelie Kroes, Vice-President of the European Commission, in charge of the Digital agenda, for her Distinguished Patronage of LREC 2012.

Organizing LREC 2012 under the auspices of these distinguish patrons is an important sign, for us who manage signs, symbols and semantics, regarding the importance conferred to languages, multilingualism, information technologies and all related fields.

These issues are at the heart of EU Digital Agenda, an Agenda that should consider Language Technologies as an essential path to pave the way to automating not only human-machine interactions, human access to information but also human-human communications, across languages and across cultures.

After having organized LREC in Marrakech and Malta, two representatives of Semitic languages (Arabic, Maltese), we are this time in a city that played one of the most noticeable roles in forging Europe, parts of Asia and Africa history and geopolitics, as well as languages, with its own language family, the Turkic languages. After a number of centuries, during which Turkish shared many aspects with Arabic and Persian including a writing system, the foundation of the republic of Turkey came with the script reform (shifting from "Arabic" characters to Latin ones) and the foundation of the Turkish Language Association in 1932 under the patronage of Mustapha Kemal Ataturk himself. The association revived so many Turkic terms and came out with so many neologisms to establish the modern lan-

guage. This experience, event if not unique in mankind history, is an important process for us, Language scientists and engineers.

At ELRA, we are very happy to carry out and support activities that help all languages to have access to resources essential for their move forward for a bright future and in particular for ensuring access to the digital world and reducing the digital divide.

We are very proud to organize this 8th LREC in that context and for that purpose: to offer our Community the forum it needs, where all players can meet and discuss hot issues related to language resources, technology evaluation, and language sciences.

With more than a thousand participants attending each LREC since 2008, we feel confident that such event where players from Academia and Industry can meet, where new comers, students and junior researchers can find background knowledge and where researchers can review new theories and trends.

With more than 1100 registered participants, more than 30 specialized workshops, about 10 tutorials, almost 700 papers at the main conference, we feel that the achievement is worth the effort we dedicate to make it happen. Boosted by this vitality and energy of our field, ELRA is moving forward with new objectives and new services to anticipate the community expectations in its challenging task to bring in more supporting tools and automations, to overcome the language and cultural barriers, and help humans enjoy the multilingualism, multiculturalism of the global world of today and tomorrow.

Over the last 17th years (1995-2012), ELRA, driven by its members' instructions, requirements, expectations, has established a number of activities to serve them. LREC is "only" and (probably) the (most) visible aspect of such services.

As many of you know, the core activity of ELRA has been and continues to be identification of valuable Language Resources, useful for research, development and evaluation of Language Technologies. Such identification, followed by a time consuming process of negotiating distribution conditions and clearing all legal issues, led to the constitution of the ELRA catalogue of over 1000 language resources and evaluation packages.

In order to help enrich such catalogue, ELRA initiated an identification process to collect and compile data on all existing resources, worldwide, to ensure that such information is shared within the community. This is our Universal Catalogue (UC). UC comprises all identified resources and a priority list is drawn before to launch the negotiations with right holders on sharing and distributing them.

To supplement this, another initiative was launched by ELRA at LREC'2010, the LRE'Map (a Language Resources and Evaluation map). LRE'Map allows each LREC author to describe resources used in his/her work. More than 1200 LR descriptions have been collected at this LREC. LRE'Map feature is now exploited by other conferences and we hope it will become a common feature to all Language Technology events (www.resourcesbook.eu). Such map contributes to spreading and sharing knowledge about LRs.

It is clear that such repositories and resources, along with fair, easy to use, and trustable legal conditions played a role in deployment of Languages Technology applications.

Since 2010, as partially reported on at LREC 2010, ELRA, through its operational body ELDA, is taking part to the META-NET Network of Excellence (Technologies for the Multilingual European Information Society). The main objective is to move forward and extend existing distribution and sharing mechanisms within a new paradigm. For this purpose, the consortium focuses on "Building an Open Resource Infrastructure", for sharing language resources and tools, referred to as META-SHARE.

META-SHARE aims to be "*a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources.*" (cf. http://www.meta-net.eu/meta-share).

One of the essential tasks of the project is related to the metadata issues with respect to the description of LRs. Work has being carried out for the specification of a metadata schema which builds upon available schemas e.g. ELRA, knowledge and expertise and provides a unified schema capable of handling the requirements of the community. These requirements comprise both the description of Language Resources and that of tools or technologies. A large number of Language Technology organizations have been debating the harmonization of such descriptions. In addition, this work aims to consider new modalities such as video and image (for e.g. sign languages, multi-sensor or multi-modal data, etc.). This work on metadata is now mature enough to be considered for standardization. More than 50 players have adopted it and many tools (metadata editor, converters from existing schemas, etc.) are made widely available.

A related issue on which ELRA and a large number of Language Technologies organization have been debating is the harmonization of the identification of LRs. A consensus seems to emerge regarding the set-up of a small executive committee, steered by a commission representing all key players in the field, data centers (ELRA, LDC, Alagin,/GSK, C-LDC,…), and the stack holders (ACL, IAMT, ISCA,…), to assign each LR an International Standard Language Resource Number (ISLRN), independently of whether the LR is accessible on Internet, Intranet, available or not, etc… whether it has a DOI, a local PId, etc. Such ISLRN should guarantee that all LR usable within our field get a unique identifier that can be used to distinguish it from others.

Another important aspect, the harmonization of existing licensing schemas and the legal aspects, has been part of the discussions and in particular, ELDA focused on the commonalities between ELRA licenses and the ones promoted by Creative Commons, with the intention to harmonize such licenses under the new umbrella of META-SHARE, which was done and will be debated during this LREC at a dedicated workshop.

A version of the META-SHARE network of repositories is already available (www.meta-share.eu) and more information about it is provided in the ELRA's president message as well as at the corresponding LREC workshop, tutorial, and several accepted papers.

As indicated above, a major barrier that hinders the sharing of language resources and tools is the copyright and other IPR issues. ELRA and the META-SHARE

partners have been working hard to offer a harmonized set of licenses that cover all needs and sharing/distributing scenarios. In parallel, ELRA continues to advocate for simplifying copyright and IPR issues concerning LRs, in particular when used for research purposes. Such exception, which exists in a number of countries (e.g. section 107 of the US copyright law), deserves to be harmonized and extended to all countries. LREC offers a useful forum for debating such issue and hopefully coming up with a common declaration on this and other similar hot topics, to be pushed forward by all of us back home.

This has been a strong credo of the FLaReNet project (in which ELRA Board members and many stack holders including ELDA) took an active role. FLaReNet conclusions at its annual forums, advocated for this harmonization. It went beyond that and compiled a useful but critical roadmap, available to all (www.flarenet.eu), and drew a clear picture of the new trends and important expectations and paved the path for ELRA activities for the coming years. Its recommendation on "Language Resources for the Future - The Future of Language Resources, The Strategic Language Resource Agenda" is an essential roadmap for us.

One of the conclusions that has been thoroughly debated within the board of ELRA is the set-up of a new permanent forum, gathering all LREC attendees and all interested individuals to constitute the Language Resources and Evaluation Forum (LRE-F). We feel that it is important to identify and gather the members of this very broad community and ensure that interactive exchanges/services can be set up to help them work together. The forum is established at this LREC 2012 where the largest group of individuals that have to do with Language Resources and Evaluation are present; it is open (and not limited) to: scientists, students or professors, involved in research activities in universities, small and medium companies or international groups; decision-makers or project managers in large public institutions, etc. You have been invited to join when registering for LREC and we hope you expressed your wish to join. Those who missed that opportunity still can do so at any time through the ELRA portal. Among the services, members of the LRE-F will be offered free downloading of many resources from the ELRA Catalogue and the META-SHARE repository, access to the legal helpdesk, access to the LRE Map, the LR Library, access to LRE Wiki, etc. Members of the community will be also encouraged to join so to upload resources on the ELRA and/or ELRA-META-SHARE repository to share with other colleagues.

An additional service offered by ELRA to all its partners, is the production, customization, repurposing of Language Resources, on demand. ELRA, through the ELDA staff, is involved in LR production. Such productions comprised speech corpora, lexica, textual corpora, both monolingual and aligned / comparable multilingual ones, video and audio data, documents and many other modalities. Such activities included production from scratch as well as, repurposing of existing ones, merging of various sets, annotations, transcriptions, META-DATA labeling of existing databases, etc. ELRA carried out such production for more than 30 languages, working proudly with hundreds of local partners all over the world.

In order to turn this into efficient and cost-effective services, ELDA is part of the EC project PANACEA (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources). The project aimed at building a factory of Language Resources that "*progressively automates the stages involved in the acquisition, production, updating and maintenance of language resources*", in particular those required by MT systems. The platform will be available both as a Framework (software package) for partners to deploy and as a service offered by ELRA for specific production of resources.

ELRA is ready to assist in LR productions, at any of the needed stages.

ELRA continues to produce resources for technology evaluation and the related campaigns. We would like to stress the importance of packaging the LR and methodologies used for such purposes, to help other interested colleagues in carrying similar assessments. It is also crucial to review such resources for possible repurposing for other needs. ELRA is prepared to assist all evaluators in these tasks. More than 50 packages are already available through ELRA catalogue, most of them for free. In order to keep an efficient stream of information on this, ELRA continues to support the HLT evaluation portal (www.hlt-evaluation.org).

While preparing this message, I went back to messages of our first gathering in Granada (LREC'1998), ages ago one would think!

> *"The presently embryonic infrastructure should be reinforced, so that the same infrastructure is able to coordinate and perform, avoiding duplications, different complementary tasks: to provide and update the general repertories of linguistic data and knowledge which should be available for as many languages as possible, to produce at reasonable costs and in due time customized LR to answer specific requests of developers, to offer services the community urgently needs, information, consultation, validation, etc."* (Antonio Zampolli, Introductory message to LREC 1998, Granada)

After the set-up and consolidation of ELRA, and now with our strong commitment to boost and sustain META-SHARE, we feel these new approaches to efficient and cost effective sharing of LRs are essential milestones for our community and ELRA is very proud to play a role in this effort.

Last but not least, let me tell you a few words about our week here in Istanbul. In addition to the technical and scientific program (see more details in our LREC Chair message, herein), we have designed, with our local colleagues, a social program to make our stay enjoyable but also fruitful for establishing new relationships and networks, setting up new projects and collaborations, and above all making new friends.

We did our best to make your stay in Istanbul a very pleasant experience, we hope that both our welcome reception (Wednesday, May 23) and Gala Dinner (Friday, 25 may) will give you memories to treasure. We hope that during these events and throughout the week, we will show you some of the best Turkey has to offer.

As always, we tried to introduce novelties and new features to improve the organization of LREC.

In addition to the EU Village, a dissemination / exhibition opportunity for EU pro-

jects, we have extended this with an EU "track" of oral presentations, to offer you a full afternoon of information on the major activities supported by the EC (Thursday, May 24).

LREC 2012 will definitely close the chapter of proceedings supplied as hardcopies, CDs or USBs. We will keep the tradition to provide the participants with hardcopies of the program booklet, and the abstracts (of papers of the main conference and the workshops, material of tutorials). BUT the proceedings will only be made available and in advance, on the LREC web site, and in various format, so that you can download them on your favorite media and bring them with you. Please do that in advance, local Internet connection may not be efficient enough for all of us to do that locally.

A new experiment will be conducted this time, a tool, called MyLREC-program, will allow participants to choose their sessions (even the papers they would like to hear within a given session), design their own program and plan their days. One can print it as a PDF file or import it in one's favorite calendar. Please visit the LREC2012 pages for this. We hope this will help you navigate efficiently and friendly through all the sessions LREC is offering.

Finally, I wish to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful.

I would like first to thank our Silver Sponsors: CELI, NUANCE; our bronze sponsors: EML (The European Media Laboratory GmbH), IMMI, K-dictionaries, META-NET, and Quaero.

I would like also to thank the EC Village participants; we hope that such gathering will offer them an opportunity to foster their dissemination and hopefully discuss exploitation plans with the attendees.

I would like to thank the LREC Local Committee, chaired by Mehmed Özkan, who helped us with all logistic issues.

I would like finally to warmly thank the joint team of the two institutions that devote so much effort over months and often behind curtains to make this one week memorable: ILC-CNR in Pisa and my own team, ELDA, in Paris. These are the two LREC coordinators Sara Goggi and Hélène Mazo and the team: Victoria Arranz, Cécile Barbier, Paola Baroni, Roberto Bartolini, Riccardo Del Gratta, Francesca Frontini, Olivier Hamon, Valérie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller.

LREC is yours; we hope that each of you will achieve valuable results and accomplishments. We, ELRA and ILC-NCR staff, are at your disposal to help you get the best out of it.

Once again, welcome to Istanbul, welcome to LREC' 2012.

Khalid Choukri
ELRA / ELDA
55/57, rue Brillat Savarin
75013 Paris, France
choukri@elda.org

Message from Mehmed Özkan, Chair of the Local Organizing Committee

Language; other than being a means of communicating human thoughts and feelings in an everyday life, is also a barrier separating civilizations apart and defining borders among the nations and peoples. Through out the history, clusters formed around languages have long provided fertile yet different habitats of cultural and social islands, enabling the cultivation of the human experience and knowledge in a relatively exclusive manner. In a way this natural isolation of human clusters made the concurrent exploration of the knowledge possible, leading to a synergetic conquest of understanding the individual's own existence. For the expected synergy to surface however, the proper means of interaction must also be present. In the old times and still for many cases the best channels of cross-cultural communication have been the multilingual humans and their products. If the language barrier is a one dominant reason for the existence of fertile cultural islands then we can argue these barriers must continue their presence, despite the transformation towards "globalization". On the other hand, for the humankind to benefit from own achievements and carry on to the next generations with advancements, the islands of knowledge are better served when shared.

With the electronic age and its substages the storage and flow of knowledge reached to a point, possibly beyond the imagination of the inventors of the information technologies, and it is growing faster than ever. In the beginning the default language to represent this massive knowledge was English. Gradually came the others. Already, it is estimated less than 20% of the information in the cyberspace is in English and decreasing. Once again the languages are claiming the borders and once again we need translators, interpreters to benefit from this vast human knowledge. This time however, the tools we need are not only for understanding the knowledge in other languages, but also to coop with the massive size and enormous speed of the information we are faced with even in our own mother tongue. LREC is playing a noble role for achieving this goal by bringing the most important resources together to tackle the problem, that is the scientist and engineers committed to develop the much needed language tools and resources.

Istanbul, located in the crossroads of civilizations, continents and important seas have lived the benefits of exchanging knowledge and prospered by attracting the scientists, artists, architects, poets, philosophers and writers from all around the world throughout the history. Named as capitals of several empires and a sultanate, was latest the European Capital of Culture in 2010. She is indeed and has been the Capital of hearts for many. For some it was the crossing of the Silk Road, for some other the ultimate spice outpost, or the final destination of the Orient Express. However we perceive it, Istanbul has been a merger point of architectural details of three continents; a meeting point of religions and of courses the languages. She lived it all… Once again it is a pleasure to see Istanbul bringing the respectable scientific community who are committed to bring people and knowledge together with innovative language and speech technologies. I am honored for being your host during this prestigious event and welcome you.

Hoping your LREC 2012 experience in Istanbul will be memorable one that you will always remember with a smile …

Mehmed Özkan
Chair of the Local Committee
Bogazici University

LREC 2012 Antonio Zampolli Prize

Speech given at the Opening Ceremony by Stelios Piperidis

This year, the Antonio Zampolli Prize was awarded to:

Charles Fillmore and ***Collin Baker***

from the International Computer Science Institute (ICSI), University of California Berkeley (USA),

and

Oriental Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (Oriental COCOSDA)

Convenors: Shuichi Itahashi, Satoshi Nakamura, Chiu-yu Tseng

From the Prize statutes:

"The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation."

Just a few words about the Antonio Zampolli Prize, the prize created by the ELRA Board in order to honour our founder and first president who did so much for the field of language resources. Citing the prize articles: "The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation. In awarding the prize we are seeking to reward and encourage innovation and inventiveness in the development and use of language resources and evaluation of HLTs". At the LREC2012 conference, the Prize was awarded for the fifth time. The ELRA Board has been very happy to receive nominations made by outstanding people in the field, and we recognize there are several persons who are eligible for this prestigious prize.

LREC 2012 Oral Session Summaries

The references given in the summaries all point to papers presented in each session. For the complete references, we invite you to refer to the LREC 2012 Proceedings, which are available online: http://www.lrec-conf.org/proceedings/lrec2012/index.html

O4 - Speech Corpora

Sophie Rosset

This session on "Dialogue and Evaluation" was very interesting and offered a large panel of work conducted in this domain. Speaking of evaluation in the dialog systems field leads to different points of view. One can consider evaluating specific modules of a dialog system or evaluating and predicting the user satisfaction, or overall dialogue.

Gordon and Passonneau in their paper "An Evaluation Framework for Natural Language Understanding in Spoken Dialogue Systems" describe experiments for evaluating different approaches for natural language understanding in two different dialog systems. Their proposed evaluation framework addresses different questions such as "is the NLU approach robust to recognizer error?" or "what is the maximum level of noise an NLU can robustly accommodate?" which are important questions when developing NLU modules for spoken dialog system. The paper compares, given different word error rates, the results obtained by two completely different approaches (CFG-based grammar and SVM-based classifiers) in two completely different domains. The results are discussed given the task difficulty and the word error rate.

User satisfaction is one of the most important metrics for measuring the performance of spoken dialogue systems. Different methods have been proposed and most of them rely on objective features among them the speech recognition accuracy. Hara, Kitaoka and Takeda in their paper "Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System" propose to estimate user satisfaction through a N-gram model based on dialog acts sequences (user, system or user+system). The best results concern the simpler classification task (complete or incomplete task). It seems obvious that the history context is useful

even for the more complex task that is the dialogue classification between satisfaction level classes (5 levels).

The need to evaluate overall dialogues and not only user satisfaction or specific modules is considered more and more important, specifically when the task or the domain is not well defined. Two papers address this question. The first one ("Dialogues in Context: An Objective User-Oriented Evaluation Approach for Virtual Human Dialogue" by Robinson, Roque and Traum) proposes a descriptive coding scheme with as objective to allow qualitative evaluations of the dialogue quality itself, for example is the kind of answer, silence or utterance, appropriate given the situation. What is interesting in this paper is that the evaluation framework allows dialogue evaluation from a dialogue perspective itself and not from a system perspective.

"Evaluating Human-Machine Conversation for Appropriateness" (Webb, Benyon, Hansen and Mival) is concerned, as the previous one, by the notion of appropriateness and propose to use this notion as a measure of conversation quality. A coding scheme is proposed regarding the appropriateness of the system (7 classes) or user (5 classes) utterances. Using this scheme, the scores offer objective and subjective performance measures which allow to show improvements over the different version of the systems.

The work described in "Constructing the CODA Corpus: A Parallel Corpus of Monologues and Expository Dialogues" by Stoyanchev and Piwek is slightly more different and concern the notion of dialogue itself as opposed to a monologue. The paper presents the construction of a parallel corpus of monologues and expository dialogues where the monologues are paraphrases of the dialogues. The dialogues are annotated with dialogue acts and the monologues with rhetorical structures. Guidelines and a tool for annotation and translation are presented.

Sophie Rosset
LIMSI-CNRS
BP 133
91403 Orsay Cedex, France
rosset@limsi.fr

O5 - Crowdsourcing Special Session

Karen Fort

The special session on Crowdsourcing consisted in two parts, beginning with four presentations in the usual LREC format, followed by a discussion about crowdsourcing. Although it is difficult to assess the exact number of attendees as the session was being held in the huge Anadolu auditorium, the audience was large during the whole session and the final discussion very lively.

The presentations offered us a large sample of crowdsourcing approaches: from GWAP ("Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources") to microworking for very different tasks ("Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing" and "Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization"), through a series of experiments testing the different approaches ("Experiences in Resource Generation for Machine Translation through Crowdsourcing").

If the ethical issues concerning Amazon Mechanical Turk were raised during the questions and the discussion that followed, the latter was more focused on GWAPs (Games with a purpose), the volunteers and the quality they produce. According to most accounts given during the discussion, crowdsourcing is more about finding (ou producing) experts in the crowd than using non-experts.

Karen Fort
INIST - CNRS / LIPN
2 allée de Brabois
54500 Vandoeuvre-lès-Nancy, France
karen.fort@inist.fr

O35 - Word Sense Annotation and Disambiguation

Yoshihiko Hayashi

In session O35 (Word Sense Annotation and Disambiguation), three papers were presented. Two of them were associated with the MASC (Manually Annotated Sub-Corpus) word sense corpus.

The first paper entitled "The MASC Word Sense Corpus" was presented by Nancy Ide, in which she described the MASC sense-tagged sentence corpus after a comprehensive introduction of the MASC, which is a 500,000 word corpus of contemporary American English with manually validated annotations in several linguistic levels. In particular, she detailed the process of sense tagging in which a number of students have participated, and discussed some important issues including the inter-annotator agreement. She also touched upon the topic of crowdsourcing to acquire useful annotations more effectively which should play a role to further grow the MASC.

The second paper "Addressing Polysemy in Bilingual Lexicon Extraction from Comparable Corpora" was presented by Darja Fišer. He described a method to extract translation equivalents for polysemous nouns from English-Slovene comparable corpora. In particular it focuses on extraction of translations for senses that are not most frequent. To achieve this goal, the proposed method first invokes existing sense taggers (UKB and WordNet::SenseRelate::AllWords) to disambiguate the polysemous English headword and computes the context vectors by combining the sense-tagging results. The resulted context vector is then translated with an English-to-Slovene bilingual dictionary to find the most similar Slovene equivalents for the target

English word. Although the proposed method yielded a better result than the results by conventional methods, he further concluded that the best results were obtained when the intersection of both sense taggers results were adopted.

The third paper "Empirical Comparison of MASC Word Sense Annotations" was again associated with the MASC project, and it was presented by Rebecca J. Passonneau. This presentation particularly compared WordNet and FrameNet as inventories of sense annotation. She introduced contingency table as a device to directly compare the annotation results based on these two resources, and proposed a new measure, called Expected Jaccard Index, to quantify the degree of association. This newly introduced measure has a number of desirable properties, of which the most crucial is that it is applicable to the cases where the inter-annotator agreement cannot be achieved. She presented a number of contingency table examples and argued that a greater number of senses does not necessarily entail lower inter-annotator agreement.

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama, Toyonaka
5600043 Osaka, Japan
hayashi@lang.osaka-u.ac.jp

O38 - Named Entities

Satoshi Sato

There are several sub-fields related to Named Entities (NE), such as recognition, classification, and categorization. The first two papers in this session have focused on named entity linking (NEL), which is the next step of the named entity recognition; a detected named entity in a document should be linked to the corresponding entity stored in a database or knowledge base if the linkage exists.

Both of the two papers have reported the resource creation for the NEL task. The first paper has reported it for the KBP (Knowledge Based Population) track of the TAC (Text Analysis Conference) from 2009 to 2011. In addition to monolingual entity linking, this paper includes resource creation for cross-lingual entity linking, which was introduced in 2011 KBP. The second paper has proposed a new methodology of resource creation for cross-lingual entity linking, which uses parallel corpora and crowdsourcing.

The third paper has reported a name matching competition, called The MITRE Challenge. Its main topic is how to design and execute the competition. The fifth paper has reported another competition of named entity recognition in historical and OCRed documents.

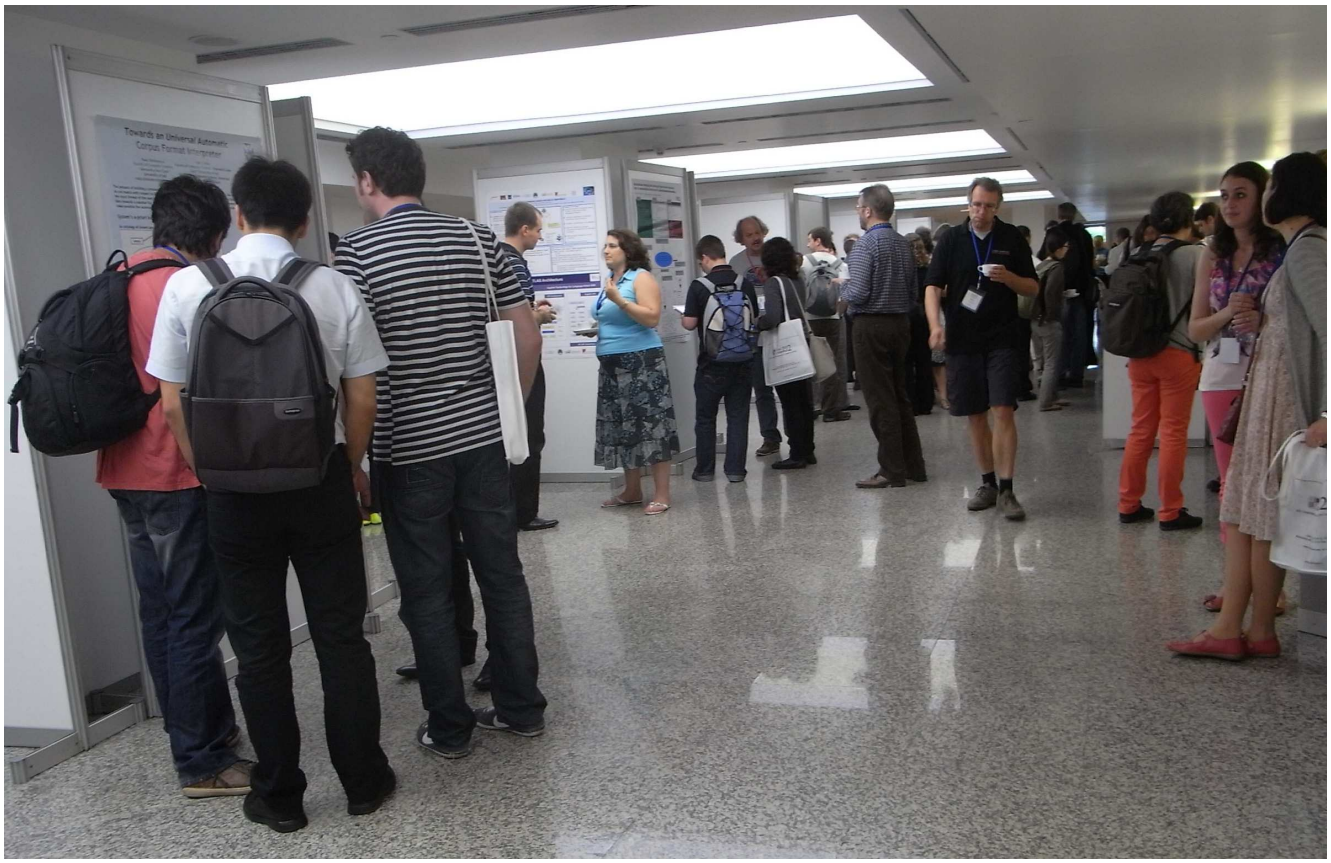
The topic of the fourth paper, "An Empirical Study of the Occurrence and Co-Occurrence of Named Entities in Natural Language Corpora", is far from the other four papers. They analyzed the pattern of occurrence and co-occurrence of NEs in several large English corpora and obtained three important observations: (1) the unique NEs grow almost linearly; (2) presence of kernel (frequent and domain-independent) NEs and peripheral (domain-specific and rare) NEs; (3) the pattern of co-occurrence of NEs shows small-world phenomenon. From these observations, they claim that "any technology designed to handle NEs should be robust for low frequency NEs."

This session confirmed that Named Entities is still a hot topic and entity linking is a crucial technology to link a document to knowledge and the real world. A breakthrough to handle low frequency NEs in language processing is expected.

Satoshi Sato
Graduate School of Engineering
Nagoya University
Chikusa-ku
464-8603 Nagoya, Japan
ssato@nuee.nagoya-u.ac.jp

LREC 2012 Poster Session Summaries

The references given in the summaries all point to papers presented in each session. For the complete references, we invite you to refer to the LREC 2012 Proceedings which are available online: http://www.lrec-conf.org/proceedings/lrec2012/index.html



P5 - Information Extraction (1)

Günter Neumann

This is a brief summary about the poster session "P5 - Information Extraction (1)" that took place on the first day of the conference from 14:45 to 16:25. I was very happy to chair this because it turned out to be a very well-attended and active session. In total, ten posters have been presented by scientists mainly coming from Europe (Belgium, Bulgaria, French, Germany, Italy, United Kingdom), but also one research group coming from China. The majority of poster presentations were about corpus work, but with a quite diverse range of topics.

For example, Wiegand et al. presented a gold standard for relation extraction in the food domain, Bank et al. proposed criteria for performing a quantitative analysis of corpora gathering and define a set of textual characteristics they consider valuable with respect to building natural language processing systems, and Zhang et al. proposed an approach for automatically extracting procedural knowledge from instructional texts. Tannier et al. focused on evolutionary aspects of event designation and presented results of a preliminary study, whereas Xia et al. focused on the important area of cross-lingual topic identification and gave details of their newly developed corpus. Also in the domain of multi-linguality, Schulz et al. presented a resource-light approach to phrase extraction for English and German which has been evaluated in the patent domain.

The effect of preprocessing and parsing for relation extraction in the domain of BioNLP was evaluated and presented by Chowdhury and Lavelli. Of particular interest of their work were the experiments on specific preprocessing techniques, for which they reported a number of interesting findings. Wang et al. presented some corpus analysis in the domain of unsupervised relation extraction. They presented a method for building a set of reference clusters of relations from a corpus that can then

be used to analyze the performance of different unsupervised RE approaches.

The final two poster presentations which I shortly introduce were more application oriented. Weise and Watrin proposed a method for the extraction of unmarked quotations in newspapers. Based on a linguistic corpus analysis they proposed 16 extraction patterns as basis for developing an extraction grammar. Initial experiments on two selected structures showed already promising accuracy rates. Last, but not least, Aleksandrov and Strapparava presented NgramQuery, a generalized query language on Google Ngram database. Implemented as tool it can be used for many applications, e.g., lexicon extraction or substitution tasks. An interesting aspect of their approach is the exploitation of phonetic and semantic similarity to increase expressibility of the actual query formulation process.

So, this should give enough flavor for reading and diving into the details of these methods and toolkits. The poster session itself was a quite lively activity. I did not counted the total number of visitors, but the whole time it was "quite crowded". I also got the impression that the program organizers did a good job when selecting these papers as posters because I got the impression that the presenters all got valuable interactive and direct feedback, something that one seldom gets during oral presentations. I personally also visited a number of other poster sessions, where I got more or less the same feeling. I think that the poster sessions at least at LREC seem to be a very suitable real-life "scientific social media".

Günter Neumann
DFKI GmbH
Language Technology Department
Saarbrücken, Germany
Günter.Neumann@dfki.de

P12 - Subjectivity: Sentiments, Emotions, Opinions

Carlo Strapparava

The field of sentiments, emotions, and opinions is becoming more and more important.

This issue fosters the creation of new linguistic resources and NLP tasks.

The present session was organized as a poster session with a total 11 presentations.

There were papers more devoted on resource construction and other on classification methodologies, about equally split between speech and written language.

While we invite to read the proceedings for the details, we can notice some interesting trends in this field. Multilinguality is becoming crucial, two papers were explicitly devoted to build multilingual emotional resources. Social media and micro-blogs are confirmed as the typical domain application for the classical opinion mining and sentiment analysis. It continues the effort to consider fine-grained and emotion classification, going beyond the simple valence classification. It is noticeable the direction to apply the classification techniques to specific practical applications.

Carlo Strapparava
FBK-Irst
Trento
Italy
strappa@fbk.eu

P38 - Subjectivity: Sentiments, Emotions, Opinions (2)

Paolo Rosso

This poster session was about subjectivity in texts, and concretely on sentiment analysis, opinion mining and emotion identification. Five have been the posters presented in this session.

Carrillo de Albornoz et al. presented SentiSense, An easily scalable concept-based affective lexicon for sentiment analysis. SentiSense is a concept-based affective lexicon especially suited for sentiment analysis-related tasks, such as polarity and intensity classification and emotion identification. In fact, SentiSense attaches emotional meanings to concepts from the WordNet lexical database, allowing to address the word ambiguity problem using one of the many WordNet-based word sense disambiguation algorithms. It consists of 5,496 words and 2,190 synsets labeled with an emotion from a set of 14 emotional categories, which are related by an antonym relationship. SentiSense has been developed semi-automatically using several semantic relations between synsets in WordNet. It is available for research purposes being endowed with a set of tools that allow users to visualize the lexicon and some statistics about the distribution of synsets and emotions in SentiSense, as well as to easily expand the lexicon.

Vázquez and Bel proposed a classification of adjectives between domain dependent and domain independent adjectives for polarity lexicons enhancements in opinionated texts. Results indicate that a majority of adjectives are domain dependent and, therefore, cannot be treated as general units. Using domain dependent lexicons allows for increasing the precision of sentiment analysis. Sundberg et al. described a graphical module based on the Qlikview software to visualize the sentiments attached to named entities mentioned in Internet forums and follow opinion changes over time.

Finally, Clematide et al. and Smedt and Daelemans presented, respectively, a multi-layered reference corpus for German sentiment analysis and a subjectivity lexicon for Dutch adjectives. The work of Cambria et al. on affective common sense knowledge acquisition for sentiment analysis was not presented.

Paolo Rosso
NLE Lab-ELiRF,
Universidad Politécnicade Valencia
Spain
prosso@dsic.upv.es

Miscellaneous Information

LREC 2012 Conference Survey Report

Following each edition of the Conference, ELRA conducts an online survey of LREC participants to collect feedback, improve the overall organization of the event and address the concerns and needs of LREC participants.

This year, 280 respondents participated in the survey. This is slightly less than in 2010 (302 respondents). The survey contained 18 questions. The majority of the responses reflected positive feedback, in particular with regard to the conference organization, and many comments, even those suggesting changes or reporting issues, congratulated and thanked LREC 2012 organization. Some responses and comments on the quality of the papers, the conference format or the need for printed material at the conference site were found very useful and will be taken into account in the next edition's organization.

Language Resources and Evaluation Journal

During the review process of abstracts submitted to LREC 2012, the reviewers are asked to assess the appropriateness of papers to be published in the LRE Journal. A number of accepted papers having met this criteria (positive review from the 3 reviewers) have been selected and their authors invited to submit an extended version of their conference paper to the LRE Journal. After a regular review process, the selected papers will be published in a special issue of the LRE Journal.

LRE Map

The new resources collected during the LREC 2012 submission process have been added to the LRE Map. The number of collected formes has reached 4260, with more than 3000 resource types covering 226 languages. The interface of the website is being redesigned and the data normalized for an enhanced browsing and access.

www.resourcebook.eu

LREC 2014

The next Language Resources and Evaluation conference will take place on May 26 to June 1, 2014 in Reykjavik (Iceland). The Harpa Conference Centre, located in the heart of the city, will host the 9th edition of the conference.

www.lrec-conf.org/lrec2014

NEW RESOURCES

Desktop/Microphone Resources

TC-STAR Resources

- ELRA-S0309 TC-STAR Spanish Baseline Female Speech Database
http://catalog.elra.info/product_info.php?products_id=1131
- ELRA-S0310 TC-STAR Spanish Baseline Male Speech Database
http://catalog.elra.info/product_info.php?products_id=1132
- ELRA-S0311 TC-STAR Bilingual Voice-Conversion Spanish Speech Database
http://catalog.elra.info/product_info.php?products_id=1133
- ELRA-S0312 TC-STAR Bilingual Voice-Conversion English Speech Database
http://catalog.elra.info/product_info.php?products_id=1134
- ELRA-S0313 TC-STAR Bilingual Expressive Speech Database
http://catalog.elra.info/product_info.php?products_id=1135

SmartKom resources

- ELRA-S0316 SmartKom Home
http://catalog.elra.info/product_info.php?products_id=1137
- ELRA-S0317 SmartKom Mobil
http://catalog.elra.info/product_info.php?products_id=1138
- ELRA-S0318 SmartKom Audio
http://catalog.elra.info/product_info.php?products_id=1139

GlobalPhone Resources

- ELRA-S0319 GlobalPhone Bulgarian
http://catalog.elra.info/product_info.php?products_id=1141
- ELRA-S0320 GlobalPhone Polish
http://catalog.elra.info/product_info.php?products_id=1142
- ELRA-S0321 GlobalPhone Thai
http://catalog.elra.info/product_info.php?products_id=1143
- ELRA-S0322 GlobalPhone Vietnamese
http://catalog.elra.info/product_info.php?products_id=1144
- ELRA-S0347 GlobalPhone Hausa
http://catalog.elra.info/product_info.php?products_id=1177

Catalan and Spanish resources for Speech Recognition

- ELRA-S0326 Catalan SpeechDat-Car database
http://catalog.elra.info/product_info.php?products_id=1148
- ELRA-S0327 Catalan Speecon database
http://catalog.elra.info/product_info.php?products_id=1149
- ELRA-S0328 Spanish EUROM.1
http://catalog.elra.info/product_info.php?products_id=1150
- ELRA-S0329 Emotional speech synthesis database
http://catalog.elra.info/product_info.php?products_id=1151
- ELRA-S0330 FESTCAT Catalan TTS baseline male speech database
http://catalog.elra.info/product_info.php?products_id=1152
- ELRA-S0331 FESTCAT Catalan TTS baseline female speech database
http://catalog.elra.info/product_info.php?products_id=1153
- ELRA-S0332 FESTCAT Catalan TTS baseline speech database - 8 speakers
http://catalog.elra.info/product_info.php?products_id=1154
- ELRA-S0333 Spanish Festival HTS models - male speech
http://catalog.elra.info/product_info.php?products_id=1155
- ELRA-S0334 Spanish Festival HTS models - female speech
http://catalog.elra.info/product_info.php?products_id=1156
- ELRA-S0335 Bilingual (Spanish-English) Speech synthesis HTS models
http://catalog.elra.info/product_info.php?products_id=1157
- ELRA-S0336 Spanish Festival voice male
http://catalog.elra.info/product_info.php?products_id=1158
- ELRA-S0337 Spanish Festival voice female
http://catalog.elra.info/product_info.php?products_id=1159

Acoustic databases for Polish

- ELRA-S0339 Acoustic database for Polish unit selection speech synthesis
http://catalog.elra.info/product_info.php?products_id=1164
- ELRA-S0342 Acoustic database for Polish concatenative speech synthesis
http://catalog.elra.info/product_info.php?products_id=1168

Portuguese Corpora

- ELRA-S0345 Spoken Portuguese Corpus
http://catalog.elra.info/product_info.php?products_id=1172
- ELRA-S0346 Fundamental Portuguese Corpus
http://catalog.elra.info/product_info.php?products_id=1173

Miscellaneous

- ELRA-S0315 A-SpeechDB
http://catalog.elra.info/product_info.php?products_id=1140
- ELRA-S0323 European Parliament Interpretation Corpus (EPIC)
http://catalog.elra.info/product_info.php?products_id=1145

Telephone Resources

LILA Project Resources

- ELRA-S0314 LILA Marathi database
http://catalog.elra.info/product_info.php?products_id=1136
- ELRA-S0344 LILA Hindi Belt database
http://catalog.elra.info/product_info.php?products_id=1170

German Speech

- ELRA-S0343 VERIF1DE
http://catalog.elra.info/product_info.php?products_id=1169

Catalan-SpeechDat

- ELRA-S0324 Catalan-SpeechDat For the Fixed Telephone Network Database
http://catalog.elra.info/product_info.php?products_id=1146
- ELRA-S0325 Catalan-SpeechDat for the Mobile Telephone Network Database
http://catalog.elra.info/product_info.php?products_id=1147

Broadcast Resources

Broadcast Resources with Named Entity Annotations

- ELRA-S0338 ESTER 2 Corpus
http://catalog.elra.info/product_info.php?products_id=1167
- ELRA-S0349 Quaero Broadcast News Extended Named Entity corpus
http://catalog.elra.info/product_info.php?products_id=1195

Written corpora

Portuguese Corpora

- ELRA-W0055 CINTIL-TreeBank
http://catalog.elra.info/product_info.php?products_id=1174
- ELRA-W0056 CINTIL-PropBank
http://catalog.elra.info/product_info.php?products_id=1176
- ELRA-W0059 LT Corpus
http://catalog.elra.info/product_info.php?products_id=1178
- ELRA-W0060 PTPARL Corpus
http://catalog.elra.info/product_info.php?products_id=1179
- ELRA-W0061 CINTIL-DependencyBank
http://catalog.elra.info/product_info.php?products_id=1180
- ELRA-W0062 CINTIL-DeepBank
http://catalog.elra.info/product_info.php?products_id=1181

PANACEA Resources

- ELRA-W0057 PANACEA English-French and English-Greek parallel corpus acquired for Environment domain
http://catalog.elra.info/product_info.php?products_id=1182
- ELRA-W0058 PANACEA English-French and English-Greek parallel corpus acquired for Labour Legislation domain
http://catalog.elra.info/product_info.php?products_id=1183
- ELRA-W0063 PANACEA Environment English monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1184
- ELRA-W0064 PANACEA Labour English monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1185
- ELRA-W0065 PANACEA Environment French monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1186
- ELRA-W0066 PANACEA Labour French monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1187
- ELRA-W0067 PANACEA Environment Greek monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1188
- ELRA-W0068 PANACEA Labour Greek monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1189
- ELRA-W0069 PANACEA Environment Italian monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1190
- ELRA-W0070 PANACEA Labour Italian monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1191
- ELRA-W0071 PANACEA Environment Spanish monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1192
- ELRA-W0072 PANACEA Labour Spanish monolingual corpus
http://catalog.elra.info/product_info.php?products_id=1193

Quaero Resource

- ELRA-W0073 Quaero Old Press Extended Named Entity corpus
http://catalog.elra.info/product_info.php?products_id=1194

Monolingual Lexicon

Arabic Dictionary

- ELRA-L0088 Arabic Morphological Dictionary
http://catalog.elra.info/product_info.php?products_id=1163

Evaluation Packages

Question-Answering and Machine Translation Evaluation Packages

- ELRA-E0039 CLEF QAST (2007-2009) - Evaluation Package
http://catalog.elra.info/product_info.php?products_id=1162

- ELRA-E0040 MEDAR Evaluation Package
http://catalog.elra.info/product_info.php?products_id=1166