The European Language Resources and Technologies Forum

Shaping the Future of the Multilingual Digital Europe

Vienna, 12 -13 February 2009

Proceedings

Edited by:
N. Calzolari, P. Baroni, N. Bel, G. Budin, K. Choukri, S. Goggi, J. Mariani,
M. Monachini, J. Odijk, S. Piperidis, V. Quochi, C. Soria, A. Toral

Istituto di Linguistica Computazionale del CNR - Pisa, ITALY

TABLE OF CONTENTS

INTRODUCTION

Nicoletta Calzolari
FLaReNet Coordinator – ILC-CNR
Flarenet_Coordination@ilc.cnr.it
www.flarenet.eu

FLaReNet – Fostering Language Resources Network – is an EC eContentPlus Thematic Network (ECP-2007-LANG-617001) whose aim is to create a shared policy and to foster a European strategy in the field of Language Resources (LRs) and Language Technologies (LTs). The growth of the field in the last years should be complemented by a common reflection and by an effort that identifies synergies and overcomes fragmentation. The consolidation of the area is a pre-condition to enhance competitiveness at EU level and worldwide.
By creating consensus among major players in the field, the mission of FLaReNet is to identify priorities as well as short, medium, and long-term strategic objectives, sustain international cooperation and provide consensual recommendations in the form of a plan of action for EC, national organisations and industry.

Work in FLaReNet is inherently collaborative. A set of Working Groups are clustered in thematic areas and carry out their activities through workshops, meetings, and via a collaborative Wiki platform. The FLaReNet **Thematic Areas** are:

- The Chart for the area of LRs and LT in its different dimensions
- Methods and models for LR building, reuse, interlinking, maintenance, sharing, distribution…
- Harmonisation of formats and standards
- Definition of evaluation and validation protocols and procedures
- Methods for the automatic construction and processing of LRs

FLaReNet is bringing together leading experts of many research institutions, companies, consortia, associations, funding agencies, public and private bodies both at European and international level. Anyone can subscribe to the FLaReNet website, joining any of the working groups and participating in their activities. This will offer the advantage of playing a role in the definition of recommendations for future actions, thus shaping the future with respect to the new challenges.

The **FLaReNet Launching Event** (Vienna, 12-13 February 2009) combined the FLaReNet themes with the i2010 objectives to address some of the technological, market and policy challenges to be faced in a multilingual digital Europe. The Forum represented an occasion to identify the grounds for future directions and strategies in the area of LRs and LTs.

The Forum was composed of a series of working sessions where leading experts were invited to present their vision on hot topics in the field of LRs and LTs. A new formula was experimented, whereby the FLaReNet Steering Committee prepared for each session a background document (here in front of each session) highlighting a set of relevant issues, and in particular a number of question to be addresses by the speakers. In all the sessions discussants (some invited in advance) and participants actively contributed to the on-going debate about priorities in the sector.
In order to elicit new ideas and perspectives from the widest possible audience, a contest was set for the best contribution and the winner was invited to give a presentation in the session on evaluation.

The final session was dedicated to a round-table on International Cooperation, mainly with non-European participants, where future policy and priorities were discussed in a global context. The aim was to initiate a strategic discussion on the utility of promoting international cooperation among various initiatives and communities around the world, within and around the field of Language Resources and Technologies.

PREFACE

Roberto Cencioni, Kimmo Rossi
European Commission, Unit INFSO-E1 –Language technologies and machine translation

FLaReNet plays an important role in the process that will define the actors, the overall direction and the practical forms of collaboration in language technologies and their "raw material", language resources. The main task of language technologies is to bridge language barriers in the global single information space, on the Web and over mobile communication devices, for spoken and written language alike. To achieve this, a community of key people need to work together and show a clear direction and priorities for the next 3-5 years.

One of the concrete tasks ahead of us is to create, for all EU languages, an open language infrastructure which allows networking of language technology professionals and their clients, as well as easy sharing of data, corpora, language resources and tools. Interoperability is a must: the common infrastructure can only succeed if the resources, tools and processes work seamlessly together, now and in the future.

The volume of multilingual information and communication is exploding in the Web. Sharing, collaboration and networking flourish – interactions are more and more instantaneous. This requires more automation: translations are needed on the fly, machine translation systems need to be set up and trained overnight, language resources need to be acquired and annotated automatically, with minimal human intervention.

The new communication and collaboration paradigms create excitement but also confusion. Language technology is a mature field, but the trusted and proven recipes may not work any more. We need new solutions and new partnerships, while securing the basic acquired knowledge base. FLaReNet will have the challenging task to create this network of people, to formulate strategies and to stimulate action in a context that is constantly changing. The demand for cross-lingual technologies is pressing, the expectations are high, and at the same time, the field is suffering from fragmentation, lack of vision and direction. In 2009 citizens will elect a new European Parliament and a new Commission will be nominated. We will deal with decision-makers that do not know us nor our business. This makes it important that we think clearly and express our ideas even more clearly.

In terms of organization, participation and stimulating debate, this two-day forum has been a big success. Now we need to reach out to the public, the policymakers and the business community – not only the academic world. FLaReNet does not have the resources to implement alone the necessary language infrastructure. Reports, meetings, events and contacts are the primary tools to achieve the ambitious goals. The success of FLaReNet relies greatly on simple things such as concise, reader-friendly reports that convey the message at first reading. All FLaReNet partners – but the coordinator in particular – have a crucial role in ensuring that all the communication matches the success of this forum.

An impact assessment has recently been completed on Language & Interaction technology actions funded by the European Commission in 1999-2005. The findings indicate that a lot of work needs to be done especially in three areas: policy, standards and outreach, especially towards the business, markets and end users. FLaReNet is an important instrument in our common effort to address these challenges.

PROGRAMME

Thursday 12th February 2009

Opening Session – 10:00 - 11:00
Chair: Nicoletta Calzolari

Roberto Cencioni (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *Head of Unit*)
Walther Lichem (*Former Ambassador of the Republic of Austria*)
Nicoletta Calzolari (ILC-CNR, IT / *FLaReNet Coordinator*)
Gerhard Budin (Universität Wien, A / *FLaReNet Local Host*)

S1. Broadening the Coverage, Addressing the Gaps – 11:30 - 13:30
Chair: Joseph Mariani - *Rapporteur*: Khalid Choukri

Introduction by the Chair

Talks:
Steven Krauwer (Universiteit Utrecht, NL) & Khalid Choukri (ELDA, FR), *"Coverage & BLARKS"*
Christopher Cieri (University of Pennsylvania - LDC, USA), *"Practical Considerations in Resource Creation Tied to Human Language Technology Development"*
Justus Roux (University of Stellenbosch, South Africa), *"An African Perspective on Language Resources and Technologies"*
Dafydd Gibbon (Universität Bielefeld, DE), *"Coverage of What? – Gaps in What? On De-globalising Human Language Resources"*
Asunción Moreno (Universitat Politècnica de Catalunya, SP), *"Shared Language Resources Production"*
Pierre Zweigenbaum (LIMSI-CNRS, FR), *,"A Dynamic View of Comparable and Specialized Corpora"*
Nick Campbell (Trinity College Dublin, IRL & NIST, JP), *"Technology for Processing Non-verbal Information in Speech"*

Discussants
Adam Przepiórkowski (Polish Academy of Sciences - ICS, PL)
Marko Tadić (University of Zagreb - FHSS - DL, HR)
Kepa Sarasola Gabiola (University of the Basque Country - IXA Group, SP)
Folkert de Vriend (Nederlandse Taalunie, NL-BE)

S2. Automatic and Innovative Means of Acquisition, Annotation, Indexing – 14:30 -16:30
Chair: Stelios Piperidis - *Rapporteur*: Núria Bel

Introduction by the Chair/Rapporteur

Talks:
Jun'ichi Tsujii (University of Manchester - NacTeM, UK), *"Richly Annotated Corpora and Their Inter-operability"*
Yorick Wilks (University of Sheffield, UK), *"Dialogue corpora remain a problem."*
Gary Strong (Johns Hopkins University - HLT Center of Excellence, USA), *"Trends in Language Resources and New Work in ASR Data Labeling"*
Dan Ioan Tufiş (RACAI, RO), *"Going for a Hunt? Don't Forget the Bullets!"*
Anna Korhonen (University of Cambridge, UK), *"Automatic Lexical Acquisition - Bridging Research and Practice"*
Gregory Grefenstette (Exalead, FR), *"The Democratisation of Language Resources"*
Marta Sabou (Open University, UK), *"Web3.0 and Language Resources"*
Iryna Gurevych (Technische Universität Darmstadt - UKP Lab, DE), *"Exploiting Croudsourced Language Resources for Natural Language Processing: 'Wikabularies' and the Like"*

Discussants
Kiril Simov (LML-IPP-BAS, BG)
Sophia Ananiadou (University of Manchester - NacTeM, UK)
Guy De Pauw (University of Antwerp, BE)

S3. Evaluation and Validation – 16:45 -18:30

Chair: Jan Odijk - *Rapporteur:* Joseph Mariani

Introduction by the Chair/Rapporteur

Talks:
Henk van den Heuvel (Radboud University Nijmegen, NL), *"The 'Standard Deviation' of LR Quality"*
Florian Schiel (University of Munich - BAS, DE), *"Towards More Effective LR Validation"*
Carol Peters (ISTI-CNR, IT), *"Evaluation of Technology for Multilingual Information Access: the Next Step"*
Bente Maegaard (University of Copenhagen - CST, DK), *"Can Evaluation Be Application-Independent?"*
Edouard Geoffrois (DGA, FR), *"Language Technology Evaluation: which Funding Strategy?"*
Bernardo Magnini (FBK, IT), *"Toward an Integrated Evaluation Framework"*
Patrick Paroubek (LIMSI-CNRS, FR), *"Evaluation: a Paradigm that Produces High Quality Language Resources"*
Harald Hőge (SVOX Deutschland GmbH, DE), *"A Proposal to Launch a Support Centre for 'Remote' Evaluation and Development of Language Technologies"*
Cristina Vertan (Universität Hamburg, DE), *"Evaluation of HLT-Tools for Less Spoken Languages"*

Discussants
Djamel Mostefa (ELDA, FR)
Nelleke Oostdijk (Radboud University Nijmegen - DL, NL)
Luisa Bentivogli (FBK, IT)

Friday 13th February 2009

S4. Interoperability and Standards – 9:00 - 10:45

Chair: James Pustejovsky - *Rapporteur:* Nancy Ide

Introduction by the Chair/Rapporteur

Talks:
James Pustejovsky (Brandeis University - DCS, USA) & Nancy Ide (Vassar College - DCS, USA), *"SILT: Towards Sustainable Interoperability for Language Technology"*
Eric Nyberg (Carnegie Mellon University, USA), *"Interoperability, Standards and Open Advancement"*
Peter Wittenburg (MPG, NL), *"Is the LRT Field Mature Enough for Standards?"*
Edward Loper (Brandeis University, USA), *"Interoperability via Transforms"*
Key-Sun Choi (KAIST, KR), *"Ontology of Language Resource and Tools for Goal-oriented Functional Interoperability"*
Thierry Declerck (DFKI, DE), *"Towards Interoperability of Language Resources and Technologies (LRT) with Other Resources and Technologies"*

Discussants
Tomaž Erjavec (Jožef Stefan Institute, SI)
Chu-Ren Huang (Hong Kong Polytechnic University, HK)
Timo Honkela (Helsinki University of Technology - CIS, FI)
Yohei Murakami (NICT, JP)

S5. Translation, Localisation, Multilingualism – 11:00 - 12:45

Chair: Gerhard Budin - *Rapporteur:* Stelios Piperidis

Introduction by the Chair/Rapporteur

Talks:
Hans Uszkoreit (DFKI, DE), "*Language Resources and Tools for Machine Translation: Trends, Demands, Predictions*"
Marcello Federico (FBK, IT), *"Outlook for Spoken Language Translation"*
Josef van Genabith (Dublin City University - NCLT, IRL), *"Three Challenges for Localisation"*
Tony Hartley (University of Leeds, UK), *"Assessing User Satisfaction with Embedded MT"*
Josep Bonet-Heras (EC - DG Translation, LUX), *"Institutional Translators and LRT"*
Alexandros Poulis (EP - DG Translation - IT Support Unit, LUX), *"Language Technology in the European Parliament's Directorate General for Translation: Facts, Problems and Visions"*
Andrew Joscelyne (TAUS, FR), *" 'Cloud Sourcing' for the Translation Industry"*

Discussants
Frank Van Eynde (Katholieke Universiteit Leuven - CCL, NL)
Harold Somers (Dublin City University - SC, IRL)

S6. Enhancing Market Places/Models for Lrs: New Challenges, New Services – 13:45 - 15:15

Chair: Khalid Choukri - *Rapporteur:* Jan Odijk

Introduction by the *Chair/Rapporteur*

Talks:

Gregor Thurmair (Linguatec, DE), *"No Resources Without Applications"*
Gianni Lazzari (PERVOICE S.p.A., IT), *"Buy a License or Pay for Service?"*
Gudrun Magnusdóttir (ESTeam, SE), *"Cheap or Expensive - What Works?"*
Gábor Prószéky (MorphoLogic, HU), *"Enhancing HLT Market with Cooperative Services"*
Jimmy Kunzmann (European Media Laboratory GmbH, DE), *"Speech-to-Text Solutions for the European Market: a SME View to Language Scalability"*

Discussants

Bob Boelhouwer (Instituut voor Nederlandse Lexicologie, NL)
Martine Garnier-Rizet (VECSYS, FR & IMMI-CNRS, FR)
Margaretha Mazura (European Multimedia Forum, BE)

Closing Session – 15:15 - 16:30

Chair: Nicoletta Calzolari

FLaReNet Sessions Rapporteurs

S1. Khalid Choukri (ELDA, FR)
S2: Núria Bel (Universitat Pompeu Fabra, SP)
S3. Joseph Mariani (LIMSI/IMMI-CNRS, FR)
S4. Nancy Ide (Vassar College - DCS, USA)
S5. Stelios Piperidis (ILSP / “Athena” R. C., GR)
S6. Jan Odijk (Universiteit Utrecht, NL)

Nicoletta Calzolari (ILC-CNR, IT)
Kimmo Rossi (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *FLaReNet Project Officer*)
Roberto Cencioni (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *Head of Unit*)

International Cooperation Round Table – 16:30-18:30

Chair: Nicoletta Calzolari

Participants

Nancy Ide (Vassar College - DCS, USA)
James Pustejovsky (Brandeis University - DCS, USA)
Gary Strong (Johns Hopkins University - HLT Center of Excellence, USA)
Jun'ichi Tsujii (University of Manchester - NacTeM, UK)
Christopher Cieri (University of Pennsylvania - LDC, USA)
Branimir Boguraev (IBM Research, USA)
Key-Sun Choi (KAIST, KR)
Nick Campbell (Trinity College Dublin, IRL & NIST, JP)
Eric Nyberg (Carnegie Mellon University, USA)
Kiyotaka Uchimoto (NICT, JP)
Chu-Ren Huang (Hong Kong Polytechnic University, HK)
Margaretha Mazura (European Multimedia Forum, BE)
Justus Roux (University of Stellenbosch, S. AFRICA)
Hans Uszkoreit (DFKI, DE)
Yohei Murakami (NICT, JP)

European Commission - DG Information Society & Media - Unit INFSO.E1 - LTs & MT:
Roberto Cencioni (*Head of Unit*)
Kimmo Rossi (*FLaReNet Project Officer*)

FLaReNet Steering Committee: Nicoletta Calzolari (ILC-CNR, IT); Khalid Choukri (ELDA, FR); Stelios Piperidis (ILSP / “Athena” R. C., GR); Gerhard Budin (Universität Wien, AT); Jan Odijk (Universiteit Utrecht, NL); Núria Bel (Universitat Pompeu Fabra, SP); Joseph Mariani (LIMSI/IMMI-CNRS, FR)

SESSIONS

Opening Session

The Societal Significance of Multilingualism

Walther Lichem - Former Ambassador of the Republic of Austria

Thank you for your kind invitation to address this prominent audience. I am certainly not an expert in your fields, an outsider, a stranger to your agenda and yet diplomats are known to speak about almost everything, most of all about subjects they have little idea about.

Yet, in defence of my presence here let me say that the outsider, the other, has become a key partner in addressing the challenges of our time. The prefix "inter" is there in our discourse, be it inter-disciplinary, inter-sectoral, inter-national approaches to understanding and policies. Allow me therefore to make some comments on a dimension of our international agenda which is of growing if not already central significance - the societal dimension of peace, security, sustainable development. And I will try to explore briefly the significance of multilingualism for this broader agenda of societal development.

What is "societal development"? The term "societal" is to be distinguished from the more traditional term "social". While "social" refers to the various dimensions of the productive capacities of the human being and of communities - health, age, education, poverty, employment, hunger etc. -, "societal" refers to the relational capacities of a citizen and of a community - capacity for plurality, acceptance and affirmation of the value of otherness, ability to relativize one's own identity, values and visions, capacity for cross-identification and, on this basis, for solidarity. "Societal" includes the capacity for understanding the common good and to articulate it in the shared public space. "Societal" provides the capacity for change, i.e. for development and for a vision of the future with change.

The societal history of the past century, and in particular over the past decades, has been marked by fundamental changes in the composition of communities, their related identities, with regard to relations among human beings and between citizens and public authorities. Socio-economic and technological development brought about an enhanced mobility of goods, services and of people, thus creating new patterns of "trans-local" relatedness. Rural-urban migrations and inter-regional movement of hundreds of millions of people persons have created a situation where "otherness" has moved from being the essence of enmity to providing the key to partnerships in our coping with the challenges of globality. I use the term "otherness" as simply referring to the state or fact of being different or distinct. Under today's conditions of regional and global integration the capacity to relate to ethnic, political, religious, social, linguistic, cultural otherness has become fundamental for achieving the security and peace as well as the development agendas.

The "trans-localisation", just to use another term for the tainted term of "globalisation", provides us with new opportunities for the interaction and integration of up to now segregate identities. These developments are accompanied by processes of societal horizontalisation. Vertical patterns of command and obedience of feudal, authoritarian, fascist regimes are ever more replaced by citizens living in conditions of self-determined identities, values and relatednesses.

As the history of the 19th and the 20th century has shown, the transition from vertically structured societies to horizontal interaction in democratic systems poses special challenges to the capacity of each citizen to define his/her identity and to relate to otherness. The concept of the single-identity society of the nation-state – supposedly sharing one language, ethnicity, religion and history - has proven to be a myth missing the fundamental societal reality of identity plurality. The new societal reality in Europe – e.g. 31,5 % of Viennese are not Austrian-born coming from close to 180 different countries of the world. In parts of Europe Islam has become the second largest religious community. In Barcelona, Spain, close to 400 different languages are spoken in homes.

"Trans-localisation" has not only enhanced the plurality of identities in our local communities but has also expanded the patterns of belonging. Information and communication technology has added new relational dimensions and capacities in our societies putting the issue of identity on to all levels of our agendas, local, national, regional/European and global. "Tell me with whom you talk and I tell you who you are!". Yet, as we communicate with the plurality of communities to which we have a sense of belonging we attain what we could call a pluri-identity personality with talking, i.e. language and the values contained in each language, becoming a central element of our own identity development.

The Slovenes in the Southern Austrian province of Carinthia have understood this ages ago when bilingual living and relating revealed their inherent duality of identities. “Koliko jezik govoris tolikokrat ti si clovek” – “As many languages you speak as often you are a human being”.

The encounters and interactions with otherness have not been easily coped with in many societies. The nation-state myth contradicted the societal reality not only in Central and Eastern Europe but also in the post-colonial countries in Africa, Asia and in Latin America. It was therefore no coincidence that the biggest office in the Secretariat of the League of Nations was the office for minorities because once the Versailles- and St. Germain-peace treaties were implemented it became evident that in every country there were other linguistic, ethnic, religious identities living in what was supposed to be single-identity nation states.

In fact, the political, peace and security agendas in the world today are marked by the challenges of plurality and often by related processes of disintegration, societal fragmentation, exclusion, marginalisation and resulting humiliation.

Wars and violence are today not any more the expression of a clash of state sovereignties but occur to more than 90 % within states, within societies, with 95 % of the victims being citizens, 85 % of them women and children.

How can we address these new challenges of societal disintegration and violence? How can we affirm the rich plurality of the linguistic, ethnic, religious and cultural identities in our societies? How can we enhance the capacities in our societies for otherness as we live the integration of diversity at local, regional and global levels? Can we build societal cohesion without falling into the trap of the “peace of sameness” or of relapsing into the “peace of authoritarian verticality” which ignores otherness and imposes single-identity societal structures by force with one language as the only official means of communication?

Recent discourse at the United Nations in New York and in Geneva has looked at the concept of “societal development” as an answer to the threats of violence and as a response to the increasing “gating” our societies.

Societal capacities in different societies have so far been largely seen as a given and not subject to policies and programmes of development. Yet as we become in multidimensional ways pluri-identity societies the capacity for otherness in our societies assumes a fundamental importance. And languages, both in their dimension of societal pluri-lingualism and of individual multi-lingualism may make an important contribution in this process.

Linguistic diversity is often seen as a dividing element in society accompanied by processes of discrimination, marginalisation and exclusion. Societal dominance often has its linguistic dimension in limiting local languages and by not admitting them into public space. The process of regional international integration logically also has its linguistic dimension. The cultural richness of pluri-linguistic societies is often lost in processes assimilation into open yet ultimately single-identity and mono-linguistic societies.

Yet languages are not only a basic dimension of identity and values as well as of otherness and diversity but also a key tool in affirming identity pluralism and in achieving societal development. It is languages which harbour the cultural richness and the historical roots of the diverse identities.

Ultimately languages are the key tool of accessing and internalising otherness. As pluri-lingualism reflects the pluri-identity society multilingualism is the key element of the pluri-identity personality. If the challenge of the 21st century is to capacitate the individuals and societies for living with a plurality of identities multilingualism will be a central element in this process. Multilingualism extended into the new means of communication, interaction, community and identity is providing the basic societal capacity for otherness and to “our common future”.

S1 - Broadening the Coverage, Addressing the Gaps

Chair: Joseph Mariani - Rapporteur: Khalid Choukri

Introduction

HLT should ensure a large coverage of languages and of the major economic/social/cultural sectors, through the supply of numerous applications and technologies which should be fed with the necessary language resources (LRs, multimedia, multilingual, multimodal).

A thorough analysis of the existing resources, technology components, etc. should be carried out along various dimensions:

1. Languages;
2. Sectors of human activities: e.g. e-services (e-learning, e-government, e-tourism, etc.), mass and specialised services, audiovisual and Internet communications, information production and access;
3. Technologies & Applications: Machine Translation, Human-Machine Interactions & Dialogue, Human-Human Communications, ML Information retrieval, access, summarisation, Subtitling, Audio-visual transcriptions and indexing, etc.;
4. Modalities: text processing, speech & acoustics, sign languages, visual/video input/output,biometrics, combination of various modalities.

Discussion, Objectives, FLaReNet Claims

A large number of the above mentioned sectors may benefit from language technologies (LT) if awareness is conducted more aggressively. But a number of such sectors lack the right applications that can be supplied by current state-of-the-art technology (e.g. broadcast news transcriptions).

A number of gaps can be identified on the various dimensions:

- If we look e.g. at trends like Statistical Machine Translation (SMT), in particular for the EU official languages (23 languages and 506 language pairs) and some “interesting/lucrative ones” (Chinese, Arabic): SMT is mostly based on European Union Jargon (JRC-acquis) and some technical manuals (Microsoft & Linux OS, etc.);
- the same comments may apply to Speech-to-Text transcription (main focus on Broadcast news of a few languages): what about other languages, other domains (conversational speech), what should be the size and content; etc.;
- and in general this applies to all LTs which can get benefit in their development from statistical approaches.

The objectives of this thread of discussion are to:

- Identify gaps along the language, application, domain, sector and modality dimensions;
- Devise means and strategies to support the development of missing LRs, especially for less developed countries and for regions, taking into account the general ecosystem (EC, Member-States, Regional governments, etc.);
- Assess/reassess BLARK/ELARK, and tailor them to current application needs, domains, multilingual and multicultural landscape;
- Suggest short and medium term actions: well-structured objectives, well coordinated tasks assigned to identified parties of excellent reputation, evaluation and monitoring of progress.

Questions

Monitoring the landscape, identifying the gaps

1.1. What are the gaps in our “scientific knowledge” relevant to the production of missing blocks?
1.2. Where are the major gaps: gaps for application, lack of technology components, lack of LRs (language & domain)?
1.3. What criteria should be considered for defining and prioritizing actions to address such gaps?
1.4. How can we identify new sectors for deploying LRs & LTs?

Extending and updating BLARK/ELARK

1.5. Starting from the definition of BLARK/ELARK, can we redefine and update these notions according to the current landscape?
1.6. Do we need to establish current “baselines” per language, per technology, etc. with a clear picture of important barriers and threats?
1.7. What are the needs /requirements /issues to tackle in order to ensure an accurate and efficient deployment of technologies for a given set of languages and domains?

Missing LRs

1.8. What are the needs /requirements /issues to tackle in order to ensure a fast prototyping for a given language and domain?
1.9. How can we promote and accelerate the extension of work conducted on one, or a small set of languages to a larger set?
1.10. How can the development of missing LRs be supported?

Suggesting directions of action

1.11. How can we identify/promote applications/technologies of “greatest exposure”? (Multilingualism?)
1.12. How can we identify the sectors than can be “early/today” adopters and how can use them as window-dressing for HLT (high exposure)?
1.13. How will we know we are making progress on addressing these gaps? (Program monitoring? Program evaluation?)
1.14. How to improve management (enforce?) of LR sharing & distribution (also from old projects)?
1.15. Which is an appropriate legal framework to foster the deployment of technologies and the successful sharing of LRs?
1.16. How enhance coordination of LRs collection between all involved agencies and ensure efficiency (e.g. interoperability)?

Coverage & BLARKS[1]

Steven Krauwer (Utrecht University / CLARIN)

1. **What it is:** The BLARK (Basic Language Resources Kit) is intended to provide a definition of the set of language resources (in the broadest possible sense) necessary to do any education and pre-competitive language and speech technology research at all.
2. **What to use it for:** It can be used as a measure of the degree to which a language is technologically covered and as an agenda for what has to be accomplished in order to create a complete collection of essential resources.
3. **Why it is dynamic:** As technology evolves the concept of what counts as an essential set of resources evolves as well, i.e. the BLARK is dynamic and has to be reconsidered and rethought with regular intervals
4. **What not to use it for**: The BLARK only makes sense for languages for which one might want to develop language and speech technologies; activities such as language documentation fall outside the scope of the BLARK concept and may require completely different types of technological support
5. **Possible purposes:** The purpose of the BLARK can be manifold, ranging from protecting a language to developing commercial products, and the notion of basic requirements may vary accordingly
6. **Different levels:** It is recommendable to distinguish BLARK levels, such as e.g.
    a. Entry level BLARKettes for languages with virtually no technological support, and mainly aimed at training and education of language and speech technology researchers
    b. Standard BLARK, serving education and pre-competitive research
    c. Extended BLARK, serving advanced research end commercial development
7. **Necessary actions:** What needs to be done now is to collect what has been done in connection with the BLARK (e.g. for Dutch, Arabic, Swedish and possibly other languages) and try to arrive at
    a. A first authoritative and broadly accepted definition of the BLARK
    b. An analysis per language where we stand
    c. Mechanisms for its maintenance (regular updates of the definition, regular analysis of the state of affairs, and if possible identification of priorities based on needs)
    d. Mechanisms and funding for the creation or completion of BLARKS for the various languages

[1] Download the presentation at: http://www.flarenet.eu/sites/default/files/Krauwer_Presentation.pdf

E/BLARK as tool for Language Resources Coverage assessment, Road mapping, and Language Policy planning. Some thoughts and considerations[2]

Khalid Choukri - ELRA/ELDA

1. BLARK is more complex than what the matrices may show:

"We define the Basic Language Resource Kit (abbreviated BLARK) as the minimal set of language resources that is necessary to do any precompetitive research and education at all. The definition is in principle intended to be language independent, but as specific languages may come with different requirements, instantiations of the BLARK may vary in some respects from language to language. For Languages that may/could afford to go beyond the basic themes of research or to address real applications we suggested to refer to the necessary resource kits as Extended Language Resource Kit (abbreviated ELARK)."

It is crucial to understand that BLARK/ELARK Concepts are oversimplifications of the HLT needs and requirements and that BLARK matrices are illustrations of very serious reporting work (and documents) that should detail various issues that can hardly fit in simple matrices see e.g. what was done for Arabic within NEMLAR:
(http://www.medar.info/The_Nemlar_Project/Publications/BLARK-final_190906.pdf).

The first assumption that one should remember is that the BLARK definition should not be seen as a static object but as a concept that over time may evolve with new technologies and application areas emerging, and with new requirements in terms of resources. The instantiation of BLARK starts with the selection of a set of applications to scrutinize. This may be the most crucial part of the work as some languages do not require the same types of resources or do not require the same components. In addition to that, the BLARK attempts to represent complex needs on (almost) two dimensional space: technologies versus components and components versus language resources. The reality is more complex, esp. when we chart with respect to specific application domains, e.g. general domain dictation machine requires a different onomasticon lexicon (proper names) from the one for medical reporting.

2. What are the missing pieces

This is a hard question. According to a market analysis conducted by Bain & Company (a global business consulting firm) for ELRA, **the traded LRs out of what exist are within the 25% for speech, 65% for dictionaries, etc.** This shows that a lot of resources exist, that are not identified within data centers and whenever they are, these are not necessarily distributable. In order to improve our knowledge of what exists worldwide, ELRA, LDC and NICT are working toward a universal catalogue that would identify all existing resources all over the world and describe them with appropriate metadata to ease the access to their documentation, fulfilling part of this need.

It is clear that even for the major EU languages, working on research e.g. on speech to speech translation requires many resources that do not exit as packages (some have pieces). If we consider that it is important to work on German to Portuguese, then a lot is still missing (though some labs have developed their own parts, not shared/not available). This can be said also about meetings transcription, conversational speech recognition, etc. to focus on speech. The same can be said for written technologies. Today's SMT development is based mostly on the European institutions texts (e.g. EuroParl) but if we go to the basics, a few languages have National Corpora (e.g. the British or American National Corpus), a few have Treebanks, etc. Regarding the building blocks, many research labs are "forced" to re-develop basic tools as they can not find them easily within the community (including Pos taggers!!).

[2] Download the presentation at: http://www.flarenet.eu/sites/default/files/Choukri_Presentation.ppt

Even for the lucky languages that have very advanced HLT state of the art, evaluation kits (the LRs, metrics, methodologies for evaluation of technologies are an integral part of the BLARK) are NOT available for several key technologies; several kits have been developed in national programs e.g. Evalda, Evalita, Dutch N-Best, etc. but these are limited both in the technology coverage and languages.

3. BLARK as an instrument for Policy makers … Simplification of the roadmaps

If one assumes that BLARK is an accurate instrument to define and implement an HLT policy, it is crucial to assess what would be the costs to fulfill BLARK recommendations for a language that has little tradition in HLT and computational linguistics in general. But let us keep in mind that a large majority of languages among the 6000 spoken today are still far away from our obvious, immediate, Information Society needs. Some, like Amazigh, have fortunately managed to revive a writing system, and even push it through the standardization bodies with Unicode coding, many have managed to get some presence on Internet, but the landscape is very worrying).

If we want to compute a nominal cost of a BLARK, starting from scratch, then we should see the cost of:

- a raw monolingual text corpus, then the cost of its Pos tagging, syntactic tagging, not to mention semantics, and what about genre, timespan, etc.
- A lexicon of a reasonable size (with all features from morphology to phonetics) and may be wordnet style information or even LC-STAR information appropriate for MT needs.
- Parallel corpus of that language with (at least) English, fully aligned for SMT, bilingual term extraction, … can we achieve 50MW with translations
- Transcribed speech , at least few hundred hours
- Language resources for technology evaluation (all these technologies?)
- etc, etc…

Who can afford that?

The finances are a key element here but not the sole factor to consider, time to market (to users of LRs) could be even more critical not to be lagging behind other languages/countries.

- ➔ Can we imagine a European Language Resources Fund (or even an international LR fund) that is set up by policy makers to invest in LRs. The sharing and distribution mechanisms may be designed to boost R&D but also HLT deployment that could generate revenues that should be invested in some other resources.
- ➔ What criteria should be considered for defining and prioritizing the actions to achieve such needs:

Can data archiving and distribution centers identify accurately the needs of the whole community and convey such message? Is the membership basis of these organizations (ELRA, LDC) representative of the whole community?

4. What else can we do to address such needs

We should also consider that LRs production is happening, daily and independently of any structured/planned action (e.g. PhD students, researchers in small teams, by players in other sectors of activity like broadcast companies or newspaper publishers, etc.) and that it is crucial to advocate for cataloguing and sharing principles to ensure that such output is not lost. ELRA with LDC and NICT is working toward a universal catalogue that would identify all existing resources all over the world and describe them with appropriate metadata to ease the access to documentation.

5. Can we leave some of this to "lucrative business principles"

Can the Market play a role in fulfilling some of the BLARK/ELARK requirements, what would that lead to?

It is unlikely that Market rules would lead to the development of BLARK pieces. One may imagine that LRs, as defined within ELARK and that can be used to develop deployable applications may be developed by commercial organizations. In this context it would be considered as a competitive advantage vis-à-vis the competitors and not released publicly. It could be released publicly only if (a) costs are shared with public funding agencies (that should/could impose some distribution at "fair" market conditions or (b) the revenues such distribution may generate constitute a serious Return on Investment for the producer.

In both conditions the roadmap will not be along the same priorities.

And Roadmap of that kind can only be short/medium term; this is truer in current economic conditions.

6. Problem of synergies in LR production (or packaging) through cooperation between communities:

Given the production/packaging costs, it is crucial to ensure that cross-technology synergies are
exploited, in terms of their requirements for LRs, whenever possible. Some simple examples:
A speech database is usually transcribed and will produce a textual corpus as a side effect (even if it is of a specific genre), such corpus can be exploited by NLP to extract terms for lexica; if the corpus is annotated with Named entities it may generate a lexicon of proper names, if the corpus is Pos tagged it may add value to language modeling used in speech recognition, etc. etc.

Such synergies are possible only with "observatories" aware of ongoing activities, like LREC, or through funding agencies when projects are supported.

We feel that BLARK concept as well as its implementation could be a good instrument to boost cooperation between research groups (in particular for pre-competitive research), while ELARK could be rather an instrument for language technology planning by funding agencies (this may be left to the HLT big players with/without assessing the consequences).

7. How can LR developed today last for ever!!

How can we guarantee that the investments of today are "safe" and will serve for the longest period possible? Or what can we do to ensure that LR produced today will continue to be usable tomorrow?

We know how to render them "easily" useless tomorrow: problems of standardization (or good practices), interoperability, quality, and archiving (with adequate metadata); it is more difficult to project the "today's resources" in tomorrow's landscape and tailor them to future application needs, domains, etc.

We have no means to guarantee that R&D and Market will continue to focus on today's themes e.g. SMT/TM are rather new and no one did guess the important role aligned corpora is playing today (well may be the pioneers of corpus-based multilingual lexica development!).

Today focus is on single modalities. Major trends concern multimodal communications that very often comprise speech, text and some image/video. What about tomorrow?

8. How to cut cost in data production …

(a) We can share: the Europarl/JRC costs us as EU citizens 3€/year ☺ and is available to all;

(b) We can be more cost-effective and automate some of the processes but this requires strong cooperation between LR producers, technology developers, and other disciplines (e.g. machine learning?): How can Today's technology help reduce the cost of producing data is used iteratively to improve its own training?

9. Quality

High quality is very often a crucial issue to bootstrap a performant baseline system. After a while a different compromise may be made between high quality and large quantity.

10. How can these considerations be linked to BLARK/ELARK:

As said above BLARK/ELARK simplifies the concept of needs / requirements of HLT community and should be linked to the corresponding detailed reports. Nevertheless it is an excellent instrument to monitor the progress of LR availability and the potential for HLT R&D and application developments.

A very rich and fully filled BLARK matrix may very well be disconnected from HLT development/deployment. It is a crucial first building block and a sine qua non for HLT development… this is simply not enough…

Practical Considerations in Resource Creation Tied to Human Language Technology Development[3]

Christopher Cieri, Linguistic Data Consortium, University of Pennsylvania

Language resource creation efforts need to be informed by specific research and development goals. This paper reports on – and discusses implications related to – three separate LDC activities that can be seen as having the side effects of broadening coverage and filling some gaps in the matrix of basic language resources for the world's languages but whose original goal was something else. The cases discussed herein show how resource building efforts require and benefit from close collaboration with R&D efforts. The projects are the GALE (Global Autonomous Language Exploitation), LCTL (Less Commonly Taught Languages) and LVDID (Language, Variety and Dialect Identification) projects.

DARPA GALE program seeks to create and integrate technologies that distill structured information from multilingual speech and text for use not by information analysts but by decision makers. The principle technologies in use are transcription, translation and distillation. Current focus languages are English, Mandarin Chinese and Modern Standard Arabic. Genres include broadcast news, news talk, newswire, newsgroups and blogs. LDC supports GALE needs through a comprehensive approach to resource creation and distribution locally undertaking or outsourcing the production of all requested resources. Audio collections include 158 hours Arabic, 68 hours of Mandarin and 32 hours of English per week covering 199 different broadcast programs collected in Philadelphia and abroad. To date GALE, has recorded more than 11,000 hours of Arabic and 10,000 hours of Mandarin. Of those about 1200 hours of Arabic broadcast news, 1700 hours of Arabic broadcast conversation and 1300 hours each of Mandarin broadcast news and broadcast conversation have been transcribed. Approximately 100 hours of Arabic broadcast news plus 140 hours of broadcast conversation and 125 hours each of Mandarin broadcast news and conversation have been translated. The inclusion of the broadcast conversation genres has introduced a significant amount of dialectal speech in all three languages, into the GALE data. Given the presence of the Arabic, Chinese and English Gigaword corpora, news text collection has been de-emphasized in GALE. However, small collections continue in the three languages in order to maintain an ongoing supply of fresh text. At the same time, LDC has intensified collection of web text, meaning weblogs and newsgroups. To date, more than 6.4 million threads have been collected. Roughly 2/3 of those are in English with the reminder evenly divided between Chinese and Arabic. Of these, approximately 500,000 words of Arabic and 700,000 words of Chinese web text have been translated. Finally over 200,000 words of Arabic and 500,000 words of Chinese source text from the web and broadcast transcripts have been word aligned with their English translations by hand.

The LCTL program created language packs and technologies for each of 19 less commonly taught languages. The language packs included monolingual and parallel text, translation encoding converters, word and sentence segmenters, translation lexicons, morphological analyzers and morphologically analyzed text, POS taggers and POS tagged text, named-entity taggers and named entity tagged text, name transliterators and grammatical sketches. The languages of focus were Amazigh, Amharic, Bengali, Burmese, Chechen, Guarani, Hungarian, Kurdish, Maguindanao, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Uighur, Urdu, Uzbek, and Yoruba.

The LVDID program supports language and speaker recognition technologies by collecting conversational telephone speech and broadcast news in a number of languages and auditing the audio for speaker, language, signal and content quality. Previous efforts have collected multiple conversations from 100 or more speakers in Arabic, English, French, Mandarin, Russian and Spanish.

[3] Dowload the presentation at: http://www.flarenet.eu/sites/default/files/Cieri_Presentation.pdf

Current collection has amassed more than 12,000 hours of broadcast news spanning 75 different languages whose exact identities will not be revealed until the NIST 2009 Language Recognition evaluation is complete. However, the LVDID data sets contain 14 of the 25 languages of interest mentioned in the call for this session as well as 6 other linguistic varieties spoken by the native, non-immigrant populations of European countries.

Resource Creation Goals and the Impact on Standards and Interoperability

The original intent of these programs was not to fill gaps in a B/ELARK for one or more languages. Each project undertook language resource creation to support the development and evaluation of one or more human language technologies. This brings us to our first issue. No single project, even a multiyear international effort that coordinates contributions from diverse national governments can manage the scope of developing all of the ELARK, or even BLARK, resources even for the 23 official EU languages let alone all the languages of Europe, or the roughly 320 languages with a million or more native speakers or the roughly 6700 languages in use today. This means that a project that seeks broad coverage of languages and resource types will likely need to fill cells in its coverage matrix with found resources created by one of more prior or independent efforts not necessarily aligning well to the new project’s needs, formats or specifications. At least some resources then must be devoted to identifying, acquiring, converting and wrapping or otherwise dealing with impedance mismatches, quality or suitability issues arising from the use of found data.

Resource Creation Goal and their Impact on Language Coverage

Within the LCTL program, languages were chosen to explore a number of issues in resource creation and technology development for under-resourced languages. Some of the languages (Thai, Urdu) were chosen to exercise a resource collection paradigm in which raw text is available digitally in sufficient quantity; others (Amazigh, Guarani, Maguindanao) were chosen to force the program to deal with cases in which it certainly is not. The cluster of Indic languages (Bengali, Punjabi, Urdu) was chosen to give researchers the opportunity to experiment with bootstrapping systems from material in related languages. Amazigh, Hungarian, Pashto, Tamil, Yoruba were chosen to take advantage of existing collaborations in order to reduce costs. Finally there was a general desire to select languages that are quite different from each other and from well-resourced languages in order to maximize the generality of our methods. As a group, the LCTL languages are linguistically and geographically diverse; they include the national languages of fourteen different countries, representing eleven major language families, in Central, South and Southeast Asia, Austronesia, North, East and West Africa, the Middle East, Eastern Europe and South America. However, only Hungarian is an official language of the European Union. Furthermore, neither of the languages identified as “interesting/lucrative ones”, Chinese or Arabic, was explored in LCTL. Finally, it is worth noting that the relative success of European researchers in producing resources for EU languages makes these same language unlikely choices in non-European programs seeking to focus on under-resourced languages.

Coverage according to Technology, Language, Genre

In any resource creation project, there is a natural competition for funding resources among coverage, volume, quality and timeline. For our purposes, coverage can be defined according to resource types, languages and genres. Each of the projects described above achieves a different balance among these factors. GALE is producing a large number of large scale resources, in multiple genres for a relatively small number of languages (3) to address takes multiple technologies. LCTL tackled a much larger number of languages (19) and resource types (18) for a single genre, news text, and produced a single medium sized resource in each cell. LVDID is currently producing a single, large scale resource, broadcast audio, for a very large number of languages (75) to address a single technology, language recognition. The goals of any resource creation project will shape its balance of coverage, volume and quality in some case yielding resources of limited immediate use to a gap filling project.

The Impact of Technology Life Cycle on Data Selection and Quality Requirements

Recent experience in GALE, LCTL and LVDID has reminded us of a number of old issues and raised some new one relevant to FlareNet goals. Project such as GALE have show us that with adequate infrastructure and effort, it is ultimately possible to exceed short term demand for language resources in at least some type/language/genre combination. Furthermore, recent experience at least suggests that there is a point of diminishing returns after which the addition of undifferentiated training data has an ever shrinking impact on system performance. As data producers reach these points focus naturally turns toward increased quality and new techniques to boost system performance. For example, by carefully coupling quality control of our Arabic Treebank with parser training we have been able to accomplish a 5.1 point, absolute, improvement in parser performance using the same engine and volume of training data, test set and metrics simply by revising the data. We are also involved in "smart" data selection techniques where the rate of agreement of multiple ASR system outputs affects the audio chosen for transcription and the proportion of novel ngrams affects the text chosen for translation. It is too soon to know the impact of these new selection procedures but the GALE community believes that smart data selection will become not only useful but necessary in the final phases of the program. Decisions regarding the quantity/quality tradeoff in resource creation depend in part on the life cycle of the research underway.

Summary

In order to be effective, language resource creation efforts need to be closely tied to research and technology development activities where the impact of decisions regarding resource creation including: language coverage, genre choice, formats, specifications and the quantity/quality trade-off can be assessed according to their impact on the ultimate goal.

An African Perspective on Language Resources and Technologies[4]

Justus C Roux - Stellenbosch University Centre for Language and Speech Technology (SU_CLaST), and CTexT, North West University, Potchefstroom, South Africa

Viewing the establishment and launch of a programme such as FLaReNet with a theme such as *Shaping the Future of Multilingual Digital Europe*, at first appears to be (geographically) exclusive. However, the qualification that the expected result is to be "… a **world wide effort** to build consensus about the sharing of data and technologies for Language Resources and applications", provides the opportunity to present some personal perspectives related to the theme from an African point of view.

Although personal, this view is informed by several experiences over the last two decades within the academic sector, as well as with government involvement (in South-Africa); it is augmented by the outcomes a seminal workshop that was held in Rabat, Morocco (2008), sponsored by the Alexander von Humboldt Foundation and entitled *Human Language Technologies in Africa: Status and Prospects.*

Development of language resources

Compared to language resource development in Europe, it is clear that (with perhaps the exception of Arabic) Africa is indeed severely under resourced. There are currently clear gaps with regard to the prioritization of the development of these resources.

Speech

From a potential **speech application perspective** (implementing ASR) it is necessary to prioritize the development of colonial languages (such as English, French, and Portuguese), i.e. to develop localized versions of English (i.e. South African, Ghanaian, Nigerian, Kenyan etc), French (Gabonese, Ivorian Coast, Congolese), Portuguese (Angolan, Mozambiquan). Most inhabitants of these areas have an indigenous language as home language, and obviously the colonial language stands to be phonetically influenced, affecting automated speech recognition. In most of the mentioned countries the first line of official communication (e.g. from government institutions) will be in the colonial language and hence the need to optimize the resources to be applicable in product development. This may also hold commercial benefits for international companies.

Although some governments profess to support the development of indigenous languages (even if they are official languages as in South-Africa) this very often is nothing more than lip-service. It can be expected though that the development of applications in indigenous languages will move to the domain of the private sector where companies will see the value of developing automated systems functioning in local languages, and hence gaining an advantage over competitors. This should be an incentive for this sector to develop appropriate speech resources for these languages as well.

[4] Download the presentation at: http://www.flarenet.eu/sites/default/files/Roux_Presentation.pdf

Text

The development of text resources in many of the indigenous African languages have to take into account aspects such as language specific diacritics (e.g. for tone indication, or purely items added to 'standard' orthographies), as well as non standardized spelling rules and conventions. This obviously impedes the development of base line applications such as spelling checkers. Furthermore, text resources in African languages are more than often only available in non-electronic format – hence necessitating time consuming scanning processes.

In the context of the **FlaReNet** aims it seems as if some of these gaps may be addressed by getting members specializing in the mentioned colonial languages interested in expanding speech corpora development to include speech varieties in previous colonial countries. Furthermore it will be necessary, possibly through FlaReNet activities, to sensitize national governments on the need for resource development and the ensuing potential economic value.

Sectors of application

Without disregarding the huge potential of commercial applications, it may be more appropriate to focus on developing applications within the following sectors within the African context: education, health, and socio-economic living environment. An important factor that needs to be taken into account is related to human computer interaction, as a large group of potential users are not technologically literate (save for the use of a mobile phone). Hence, area / domain / language specific interfaces may need to be considered. Given the discrepancies in internet penetrations within African countries, some may obviously be more ready for applications than others.

Within the assumed FlaReNet activities, it seems that much could be learnt from countries in similar situations, not only in Europe, but also in, for instance India.

Technologies, applications and modalities

Country specific analyses should be made with respect to the implementation of the most appropriate technologies and applications, ranging from high level machine translation requirements to specialized human-machine dialogue interactions. I would like to argue that, given

(a) the high rate of illiteracy in many African countries,

(b) the relatively limited penetration of the Internet in African countries, but

(c) the extremely high penetration of mobile phone services across Africa,

emphasis should be laid on the development of speech based applications on mobile platforms.

Within the FlaReNet context it would be excellent if African researchers and developers could have access to European infrastructures, language resources and technologies. Simultaneously they will be in a position to share the challenges of African multilingual societies with European counterparts, who may also be involved in multilingual language and speech processing.

Coverage of what? – Gaps in what? On de-globalising human language resources[5]

Dafydd Gibbon (Universität Bielefeld, COCOSDA Convenor)

'Coverage' and 'gap'

My main concern in this contribution is with the responsibilities of the scientist and technologist in relatin to under-resourced and endangered local languages. Both 'coverage' and 'gap' are highly context-dependent and multidimensional, and we tend to associate them with our own scientific and technological interests as language and speech scientists and technologists.

The term 'coverage' may refer to whole languages as well as their lexical, grammatical or phonetic components, to under-resourced as well as to well-resourced (perhaps over-resourced) languages, to the dialects and varieties of languages of which their speakers are justly proud, to text genres and situation dependent sociolects and speech styles, as well as to the history of all these.

The term 'gap' is very different from the term 'coverage', however: it presupposes a sense of 'wholeness' and 'deficit' in relating to the extension of an existing state or the materialisation of a platonic ideal. In respect of the notions of coverage just outlined, the notion of 'gap' is perhaps inappropriate: our 'coverage' is a tiny island in a huge ocean. As scientists and technologists of language and speech, with these terms we focus on our own interests in advancing knowledge and technologies to our own benefit, and perhaps also to the benefit of humanity, within the multidimensional global information, economic, political and military society of which we are a part.

Globalisation and de-globalisation

So my first thesis in this context is a political one, but it has immediate scientific and technological consequences, since the thesis is concerned with the our responsibility as scientists and technologists:

> ***Globalisation is – for better or worse – no more than a patina on a much deeper and more complex sculpture of cultures, religions and societies, an illusion shared by elite networks in politics, economics, the military, as well as in science and technology.***

Global networks are paradoxical constructs: the nodes in the network are local, and have their own local constraints. Behind these local nodes lies the – in our perception silent – majority, with its own values which are far away from our own, or from the currently disintegrating financial snowballs, and which remain relatively untouched in daily life by the languages of current globalisation forces, whether English, Spanish, French or German, and will remain untouched by the languages of future globalisation forces, whether Chinese, Russian and Arabic. This majority has practically no part in the globalisation process, except to endure its effects: global warming, industrial pollution, false hopes of sharing in the process, and the digital divide.

So my second thesis is a complementary one:

> ***De-globalisation of global political, scientific, political, military, scientific and technological monocultures, in the sense of maintaining respect for non-global cultures and languages, is a necessary condition for the survival of humanity, and with it of the science and technology of language and speech.***

Reasons and consequences

Why should this de-globalisation be the case, and what does it really mean? The primary reason for a scientist is that the limits of human knowledge, reason and understanding will be cemented as globalisation increases. The languages of the world, which are threatened by globalisation, show a

[5] Download the presentation at: http://www.flarenet.eu/sites/default/files/Gibbon_Presentation.pdf

variety and complexity of form and function which can support an intellectual and practical trading relationship between equals:

1. Knowledge of the complex forms and functions and functions of the world's languages can help us as speech and language scientists and technologists to deepen and apply our knowledge of our own languages by challenging us to refine our theories and models and expand their coverage to very different language types. These structures – whether morphological prosody, or serial verb constructions, or click phonemes – occur in different guises in our own languages, but generally remain unrecognised as such because of our dominant highly conventionalised and standardised written national languages. Other varieties and styles are in many ways richer.
2. The 'technological gap', and its special case, the 'digital divide', is taken to be asymmetrical: the patina of global societies is taken to be on the higher ground, and the non-global societies are taken to be on the lower ground. And if technological advancement is taken as a measure, this is true: my colleagues in Nigeria are privileged in relation to most of the country's populsation but, despite the mineral oil wealth of the globalisation patina in that country, they struggle daily with unreliable power supplies, breakdown and slow speed of internet connections.

So what does this mean, in concrete terms? First, it means that we have an intellectual responsibility to extend 'coverage' to other languages, dialects and varieties, to the extent that others wish to share these with us. Second, it means that we have a practical responsibility to offer technological facilities – I deliberately avoid the word 'benefits'! - to societies *which desire them*. In both scientific and technological fields, it means specifically that we must offer – perhaps better: continue to make offers and to extend our offers – of cooperation in education, research, and development to our colleagues as equals, but in infrastructurally weaker areas of the world, and not comfortably just tie ourselves into the funding paradigms which skate over the patina of globalisation. I will call these 'global but non-global cooperations'.

Many of us do have intensive global but non-global cooperations already: in cooperation with universities and other institutions in Africa, Asia, Oceania and South America. I propose that we go further: that every research group and funding agency deliberately cultivates and actively sponsors cooperation with and support of colleagues in these areas. There are encouraging signs that this kind of cooperation is growing, and will increasingly complement traditional forms of cooperation: being invited for short stays, taking on doctoral students, organising brief visits. A very encouraging sign occurred in 2008: the staging of the first major scientific conference on language and speech in the whole of Africa, the LREC conference in Morocco, which enabled the participation of many scholars who would otherwise not be able to do so. This is a model which needs to be extended: a form of globalisation which at the same time faces the challenge of de-globalising in respect of entrenched habits and interests.

Next steps

So what are the steps to be taken? There are infrastructural steps of the kinds which I have mentioned, which can range from personal involvement to the creation of political lobbies for infrastructural strengthening. But there are practical steps for the two most important scientific and technological areas within the remit of the FlaReNet group:

1. For *written language*, development of tools for combatting the sparse data problem for our 'gaps', the legacy documents in arbitrary fonts and formats, and for enhancing future data capture, for sustainable information storage and provision.
2. For *spoken language*, development of robust, language-independent (semi-)automatic tools for the capture and annotation of speech (and multimodal) data, and the development of speech

synthesis and recognition tools for data validation, model-checking and applications for multimodal communication.

3. The development of new modes of sharing in education, research and development, and funding, which permit the equal advancement of language and speech science and technology in both global and non-global societies, and the devolution of resources and archives into distributed local responsibilities.

In several publications and lectures over the past decade I have advocated a ‘code of conduct’ for offering (and not enforcing) such cooperations, wherever wanted, in order to bridge the many gaps and divides. They must be:

Comprehensive – Efficient - State of the art – Affordable - Fair

The current re-shuffling of the world’s economies may well support the kind of de-globalisation of interest that I am referring to. We would be well-advised to pay even more attention to what is going on, in terms of its effects on our focussing on ‘coverage’ and ‘gaps’ in the language and speech sciences and technologies and respect for the domains beneath the thin patina of global societies. And perhaps the election of a gentleman with a multilingual, multicultural and multiethnic background to be the head of the most powerful force for globalisation in our time is also a sign that this kind of trend towards de-globalisation is on the increase.

Shared Language Resources Production[6]

Asuncion Moreno
Talp-UPC

Production of Language Resources is expensive and labour intensive. To resolve this problem, several attempts have been made to produce large databases within consortia of companies and universities. These consortia allow producers to share costs and work. If consortia are open to new partners, the number of produced databases increases as new partners add to the group. Currently, shared Language Resources production has become a common practice between companies.

The first consortia created to produce large databases were born in the last decade. The European commission promoted the production of language resources in most of the official European languages. During several EU calls, large, funded consortia have created a number of large databases. Examples of such projects include the *SpeechDat family* where big databases for ASR were produced for telephony (SpeechDat M, SpeechDat II, SpeechDat East, Orientel) in-car applications (SpeechDat-car) and commercial applications (Speecon).

The large databases generated in the projects belonging to the SpeechDat family are the result of a successful production model comprised of several parts:

Specifications: A set of specifications that exactly defines the linguistic contents, number of speakers, ages, recording environments, formats, and documentation.

Production: Each partner of the consortium produces a database in a given language that exactly fits the specifications.

Validation: An external organization: a) ensures that the produced database fulfils the specifications, b) provides linguistic experts that check the produced transcriptions and lexica, and c) measures audio signal quality. If a database is not positively validated, the database has to be repaired.

Exchange: All positively validated databases are exchanged between the partners.

Extending the model

The SpeechDat model, without financial support from the EU, continued in projects like SALA (I and II) where a number of companies produced databases for telephony applications in America (North America and Latin America) or LILA with the objective of producing databases in Asia. These consortia are open to new partners. The new partner must produce a new database that fits the specifications and validate it. Then, the partner is ready to exchange its database with other partners. This production model has been a very effective, and cost efficient way for small and medium sized companies to obtain large data sets in a number of languages.

The same model has been extended to the production of Lexica. Indeed, the EU funded project LC-STAR, “Production of lexica and corpora for speech to speech translation”. followed the SpeechDat family model and continues today with LC-STAR II and III with more than 20 lexica already produced.

[6] Download the presentation at: http://www.flarenet.eu/sites/default/files/Moreno_Presentation.pdf

Lexica are comparable in terms of number of words (common words and proper names), and each entry word in the lexicon contains POS, lemma, and phonetic transcription. The model has been successfully applied to European, Asian, and Arabic languages.

In speech synthesis, ECCES consortia follows the same four parts (specifications, production, validation and exchange) as SpeechDat model. ECCES is mainly comprised of research centers and Universities.

Discussion of the model

Advantages of the model are: the quality of the produced databases is comparable, databases are produced in a specified period of time, and the total cost of the project is affordable assuming the number of partners is above a given threshold.

Disadvantages of the model are: sometimes not all the partners are interested in all the produced languages; delays and withdrawels can impact internal companies planning; specifications do not fit exactly the needs of a company; and divergences can occur between criteria of producers and validation centers

In addition, there are some other considerations for shared production without external funding:
- The model has been applied to those databases where specifications exist from a funded previous project.
- The model has some, albeit limited, impact in the production of LR for new research topics.
- The model is not attractive enough for research centers and Universities.

A Dynamic View of Comparable and Specialized Corpora[7]

Pierre Zweigenbaum - LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
pz@limsi.fr

Corpora, coverage and gaps

A corpus, according to Sinclair (1996), is “a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language”. An important point of this definition is that a corpus is not a mere (large) heap of texts, but depends instead on specifications designed to meet certain

goals. Coverage and gaps can only be measured against these goals and specifications. The position I defend below is that these are moving targets, and that a dynamic approach must therefore be taken whereby suited (comparable and specialized) corpora can be designed and assessed as needed to meet evolving needs.

Comparable Corpora

Machine translation, especially statistical, needs to be trained on large numbers of examples of existing translations. Therefore, they depend on the existence of large parallel corpora, i.e., corpora made of pairs of texts and their translations. Among the most often used parallel corpora are parliament debates in multilingual settings (Hansard:

English, French, Inuktitut; Europarl: languages of the European Community; Hong Kong: Chinese, English). Multilingual Web sites provide another source of parallel texts, which can be harvested with appropriate methods (Resnik & Smith, 2003). Parallel corpora are however limited in quantity and diversity. Only a few language pairs are represented with a substantial amount of data, English often being part of the pair.

To overcome this bottleneck, methods have been proposed to exploit non-parallel, multilingual corpora, called comparable corpora (Fung, 1995; Rapp, 1995). Two corpora are comparable when although not being translations of each other, they are similar in nature: they address the same topics, have the same genre, etc. For instance, newspaper articles about international politics extending across the same time span are expected to cover shared events and therefore to express similar information in different languages; texts about diabetes in two different languages are liable to describe similar knowledge. Extracting translation correspondances from comparable corpora is more difficult than from parallel corpora. On the bright side, the potential size and variety of such corpora are much larger than those of parallel corpora, since constraints on their composition are lighter. The basic requirement is to find two monolingual text corpora with similar coverage; building independent monolingual corpora is much easier than preparing a parallel corpus (although I discuss this point below). Another important feature of comparable corpora is that they can (and should) be original texts. Translations, whatever the skills of professional translators, run the risk of containing “calques”, i.e., expressions patterned after the source language, where the use of more natural native expressions would have been possible and more appropriate. Inasmuch as comparable corpora are made of original texts, they remove this bias in the language samples they provide. These features make comparable corpora very appealing for machine translation and other multilingual tasks.

There is a need therefore for comparable corpora to alleviate the scarcity and limitations of parallel corpora. Few comparable corpora have been built up to now, mostly to test methods for extracting word translations (Zweigenbaum et al., 2008). We have seen that to be comparable, two corpora must be built according to controlled specifications: common topic, but also similarity along

[7] Download the presentation at: http://www.flarenet.eu/sites/default/files/Zweigenbaum_Presentation.pdf

other dimensions such as genre, date, etc. The quality of language resources extracted from such corpora depends on this control. It is questionable therefore whether a comparable corpus can be designed to fulfill general needs. Instead, designing targeted comparable corpora for each need seems a more appropriate strategy. The variety of needs is that of translation: translating news, parliament debates, device manuals, scientific articles or blogs are but a few examples. Building comparable corpora for some of these targets will indeed be useful, but more important is to provide methods and tools to help develop more quickly and more reliably multiple types of comparable corpora as they are needed.

Given the existence of large repositories of texts such as the Web, the main difficulty lies in the principled selection of appropriate texts from these repositories. A key point for this to happen is to produce a more precise definition of comparability, which will help design comparable corpora, along with measures of comparability, which will help assess to which extent they qualify as comparable corpora.

Specialized Corpora

Activities in specialized domains generally create and use specific terminology and more generally define sublanguages: law, medicine, car manufacturing are well-known examples. Language resources for specialized domains therefore include terminologies, possibly linked to ontologies to provide more formal support. Specialized text corpora are also a key asset: term extraction software can spot new terms in texts to help create new terminologies or maintain existing terminologies. Multilingual specialized corpora are also needed to support the acquisition of translation resources (see previous section). Accessible specialized text collections include abstracts of scientific publications (MEDLINE database in the biomedical domain), JRC Acquis (body of European Union law in 22 languages), both being evolving collections as new articles or laws are produced.

Terms are the building blocks to describe information or knowledge, whose value depends on its being up-todate. For relevant terms to be spotted, specialized corpora must therefore be up-to-date too. Static corpora would soon be outdated and mostly useless: this precludes the preparation of a warehouse of specialized corpora. Besides, specialized corpora must be built for each specialized activity: they must be selected according to a domain (e.g., medicine), but also to the intended audience (e.g., general public vs health care professionals) and possibly many more dimensions. The multiple existing terminologies in the health domain reflect the diverse needs of different activities (e.g., nursing, care, public health), tasks (diagnosis, procedures), specialties (surgery, medicine), etc. Very different corpora are needed to support such different terminologies. Foreseeing all needs for such a variety of situations is close to impossible, which adds strength to the above argument against an a priori collection of static specialized corpora.

I would therefore put forward here again the principle of on-demand selection of relevant texts from a larger, evolving collection. Corpus construction is then split into two steps: (i) construction of and/or access to a large collection of texts which should be a superset of the target corpus; and (ii) selection of the actual corpus according to specified criteria targeted to the needs at hand. This way, the target corpus can be kept up-to-date as the source collection evolves, a key to actuality and maintenance. For this to succeed, methods are needed to characterize and measure dimensions of specialization (topic, audience, addressed tasks, language level. . .). An additional constraint is to control the reliability of the source of a given document: for instance, which organization or individual backs a given text is key information for a user to trust a source and the language and terms found in its texts.

This whole section was dedicated to specialized corpora. The existence of non-specialized corpora and the notion of “general language” are sometimes debated though: what is generally considered as instances of general language, e.g., newspaper articles or novels, may be seen as just other instances of specialized corpora, being specific text genres covering certain topics. The issues discussed above might then have even larger applicability.

Access issues

Legal issues plague the development and distribution of corpora: either the included texts must be free of copyright (or have copyright statements which explicitly grant the right to copy and redistribute texts), or permission must be obtained from the copyright owners. The latter is rarely feasible with a corpus obtained from the Web because of the large number of different sources.

Privacy issues are additionally involved in domains such as clinical medicine, where the need to de-identify patient-related documents has hampered the development and sharing of specialized corpora, hence of shared language technology evaluation tasks.

References

FUNG, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In Proceedings of the 33rd ACL, pp. 236–233, Boston, Mass.

RAPP, R. (1995). Identifying word translation in non-parallel texts. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, student session, volume 1, pp. 321–322, Boston, Mass.

RESNIK, P. & SMITH, N. (2003). The Web as a parallel corpus. Computational Linguistics, 29, 349–380. Special Issue on the Web as a Corpus.

P. ZWEIGENBAUM, ÉRIC GAUSSIER & P. FUNG, Eds. (2008). Proceedings LREC Workshop Building and using comparable corpora, Marrakech, Morroco. ELRA. Available at http://www.lrec-conf.org/proceedings/lrec2008/workshops/W12_Proceedings.pdf.

Technology for Processing Non-verbal Information in Speech[8]

Nick Campbell - Trinity College Dublin, Ireland

Current speech technology is founded upon text. People don't speak text, so there is often a mismatch between the expectations of the system and the performance of its users. Talk in social interaction *of course* involves the exchange of propositional content (which *can* be expressed through text) but it also involves social networking and the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. A computer-based system that processes human speech, whether an information-providing service, a translation device, part of a robot, or entertainment system, must not only be able to process the text of that speech, but must also be able to interpret the underlying intentions, or *acts*, of the speaker who produced it. It is not enough for a machine just to know *what* a person is saying; it must also know *what that person is doing* with each utterance as part of an interactive discourse.

Tone of voice

Previous work carried out in Japan has shown that more than half of interactive speech in everyday conversations takes the form of nonverbal utterances which cannot adequately be transcribed into text. These stylised utterances as well as non-lexical *affective* speech sounds, such as laughs, feedback noises, and grunts, also carry important interpersonal information related to the states, intentions, and beliefs of the discourse participants, and to the progress of the social interaction as a whole. They constitute a small finite set of highly variable sounds in which most of the information is carried by prosody and tone-of-voice. It is this component of speech especially that makes it such a rich and expressive medium for human interaction, but this is an element of the signal that is not yet well modelled, if at all, by machine processing.

A human interlocutor intuitively interprets the nonverbal information in speech and tone-of-voice to aid in the interpretation of each utterance in context. It has been shown, for Japanese, that a machine can be programmed to perform similar interpretation of speech utterances, and currently research is being carried out to generalise and further develop these findings using speech data from other languages. While the academic goal of such research is to show that the use of nonverbal utterances in conversation is a characteristic of human speech *in general* and not limited to only one particular culture or language, the technical goal of the work is to produce devices that are *specifically adapted to interactive or conversational speech* that will enable a friendlier and more efficient speech interface for public services and entertainment.

Recognising that *social actions* are the essential component of intercourse, and that actions, rather than words are the prime units to be processed in a discourse, future speech research must specifically address the question of how new technologies can be produced which are capable of processing not only the lexical content of an utterance, but also its underlying intentions. This might be done by processing prosody & tone-of-voice.

To further the development of such speech technology, it is therefore essential to collect a representative corpus of spoken interactions wherein participants display the *full range of their daily speech strategies* and to use that material to train new modules for interactive speech processing

[8] Download the presentation at: http://www.flarenet.eu/sites/default/files/Campbell_Presentation.pdf

(whether for synthesis or recognition) that can make use of such higher-level information. However, such a corpus requires the prior development of recording techniques that are unobtrusive, and environments which are felicitous.

Discourse dynamics

There is growing international interest in multimodal interaction processing (see e.g., UC *(Universal Communication)* in Japan, AMI *(Augmented Multimodal Interaction)* in Europe, and CHIL *(Computers in the Human Interaction Loop)* in the US) and in the collection of multimodal conversational speech data, which was identified as a principal future task at the LREC (Language Resources and Evaluation Conference) last year.

Whereas traditional approaches to spoken interaction and dialogue systems have tended to assume a “ping-pong” or “push-to-talk” model, wherein either the system or the interlocuting human is active at any given time, it is becoming increasingly apparent that the dynamics of spoken interaction is an important element in itself for speech information processing, and that the typical flow of speech is fragmented and multi-faceted, rather than forming a single uninterrupted stream. This is supported by many recent findings in conversation and discourse analysis, where the definition of a “speech-turn”, or even an “utterance” is proving to be very complex.

People apparently don’t “take turns” to talk in a typical conversational interaction; rather they each contribute actively *and interactively* to the joint emergence of a “common understanding”. The apparent "no gap no overlap" alternation of spoken utterances is actually emergent from a background of continuous behavioural coordination at different levels of behavioural organization. This *interaction synchrony* is a feature yet to be incorporated in modular speech processing technology and might prove to be an important element for dialogue interface design. It should therefore be taken into consideration as a key component of corpus design.

Corpus control

Speech data will continue to be collected from a variety of sources using a variety of capture devices. Techniques will be developed to deal robustly with impoverished or "less-than-perfect" materials, and a corresponding robustness will be reflected in the technology produced as a result. Conversely, in order to derive useful and reliable components for speech information processing, we should ensure that the corpora we collect are representative of the styles and mannerisms of interactive conversational speech, so that future users of this technology will be presented with interface designs that match their (unconscious) expectations and that are able to process the full range of information that is carried by inflections of the voice and from the characteristics of timing and turn-taking.

Conclusion

As we envisage the incorporation of speech processing modules in more and more sophisticated commercial applications, including machine interpretation, robotics, games, and customer-services, a key element of the research will be to develop methods that enable the efficient collection of conversational and interactive speech data without the need for extensive or invasive recordings. Privacy considerations may prevent the use of naturally-occuring samples, so this work may require the development of both capture devices (cameras and recorders) and capture environments (equivalent to a recording studio) that encourage participants ro relax informally and maximise their range of speaking styles and formats.

S2 – Automatic and Innovative Means of Acquisition, Annotation, Indexing

Chair: *Stelios Piperidis* - *Rapporteur*: *Nùria Bel*

Introduction

Contemporary methods for language technology (LT) R&D rely on the deployment of the appropriate language resources (LRs) more than ever before. The most promising LTs, although language independent by themselves, are nonetheless inherently tied to language dependent knowledge in the form of LRs. This paradigm shift, effected in the late eighties, applies today almost to all areas of LT: from speech recognition and synthesis to technologies for converting unstructured information (textual or multimedia) to structured information by means of a range of information extraction technologies and contemporary methods for machine translation technologies development. Components and tools enabling the development of these applications rely heavily on LRs – such as lexical resources, annotated or un-annotated corpora, ontologies – depending on the learning technique adopted. At the front of multilinguality and machine translation, the success of statistical machine translation renders multilingual resources the absolute indispensable requirement.

In their turn, the use of LTs can be looked at as a source of competitive advantage, especially if they are considered as general purpose technologies that can add value to most ICT products dealing with language in whatever manifestation. But multilingual technologies are located on the production-side of the economic equation. They are intermediate products used to produce final goods and services, and therefore they are valued for what they actually do. And what LT based applications currently do is hampered by the fact that eventually they fail when they need to cover a new word, or a new domain, or a new language. An additional challenge to the robustness, coverage and performance of the tools and applications mentioned above is presented by the current language use on the various web communities, social networks, blogs and the like. Moreover, language on the web and on other information and communication platforms (radio, TV, etc.), converging today through advances in telecommunications engineering, is tightly interlinked to other media, notably images, video and sounds. The unavailability of the appropriate resources is a hindering factor for systems and application development and full deployment. It is therefore of uttermost importance to develop and deploy methods for an automatic construction, linking and repurposing of new and existing LRs which can satisfy such demand.

Discussion, Objectives, FLaReNet Claims

Automatic techniques and methods for speeding up the resources building process include the building of purpose specific tools from using the latest technologies on automation and machine learning, e.g. specifically trained web crawlers for parallel data identification and automatic multilevel aligners to ontology learners, lexical mining, etc. Issues to discuss wtr are the relationship between quantity and quality of language data and its impact on technologies, and the possible improvement achieved by the current results of automated production of LR’s. Current methods vary in performance but are they mere experiments or should be consider near to production techniques?

Besides, the rapid evolution of collaborative methods and networks is gradually designing a new paradigm of creation and development of knowledge repositories. As particular types of knowledge repositories, even language resources could take advantage of methods of collective construction of knowledge by non-specialist volunteers. This calls into place the creative thinking of radically different

modalities of resource creation and deployment, demands new technological solutions, and opens up unforeseen scenarios of resource validation. What are the experiences with such approaches so far? Are they indeed successful in terms of production quantity?
Further challenges to the LRs and LT field emerge today as requirements from developers of cognitive and robotic systems. Not infrequently, innovation in cognitive and robotics applications require new types of LRs, or sometimes a different view to the content of existing LRs (for instance, the lexicon of natural language), such that a kind of language resource re-engineering is necessary.

Questions

Drawing on the above, the following issues are to be elaborated:
2.1. What are the current methods for automatically building/linking/repurposing LRs? Are such techniques usable already on a large scale, or are they still research efforts?
2.2. Regarding the automatic creation of resources: does quantity equal quality eventually?
2.3. Are there successful stories we can learn from and build on for the future?
2.4. What is the future target of LR acquisition/annotation? Can we set priorities?
2.5. Are the existing resources suitable for the development of current applications, systems etc.?
2.6. What is missing in the current picture? What new types of resources are necessary?
2.7. Is the cost of ensuring interoperability / compatibility worth paying? How can it be quantified?
2.8. Is the interlinking between individual monolingual resources for the development of multilingual resources a viable solution? What does it entail?
2.9. While LT has been traditionally developed for processing well-formed language (text), language use on the web is largely “unregulated”: how effectively can we process today the language used on the web, in blogs, in chat rooms and other web-based forums?
2.10. Can LRs be used by other disciplines and new areas: cognitive systems and robots for instance. Can we repurpose them? What types of LRs and LT does the contemporary intelligent humanoid robot need to acquire or improve its linguistic capacity?
2.11. What does the linking between different media (language, images, video, sounds) entail?
2.12. Are LRs to be optimized in the future taking new shapes, as a result of automation processes?
2.13. On the other hand, how can the web help in delivering quality language data and annotations of them?
2.14. How can the web 2.0 and the collective intelligence be used in the production of LRs? What role can the now popular social networks play in the acquisition of language use data?
2.15. Can automatic LR production participate and contribute in the web 3.0?
2.16. What are the experiences with social/collaborative approaches so far? Are they indeed successful in terms of production quantity and quality?
2.17. How are people encouraged to participate? Why is it in their interest?
2.18. Can collaborative techniques be used to do annotation in the area of speech and language resources? For example, can we devise an entertaining web-based game that will yield large quantities of high-quality phonetic transcriptions for names, or that yield annotations to place a word in the right place in a WordNet like database?
2.19. Are there any properties of the successful approaches that should be taken into account? (e.g. in the case of picture tagging there is no absolute standard, many different tags are appropriate, and the fact that multiple independent individuals come up with the same tag for a picture is by itself proof of its usefulness as a tag for this picture).
2.20. Can collaborative web data be exploited in order to derive new types of language resources and if so, how?

Richly Annotated Corpora and Re-usability of Resources[9]

Junichi Tsujii - National Centre for Text Mining and School of Computer Science
University of Manchester, UK
Department of Computer Science, University of Tokyo, JAPAN
tsujii@is.s.u-tokyo.ac.jp

After fruitful use of syntactic tree banks for research in Natural Language Processing, the community has become interested in much richer annotation of text, including deep syntax, semantics, co-references, discourse structures, etc. Unlike shallow phrase structure, such “deep annotations” often lack theories which play the role of common minimum denominators among different research groups. As a result, several annotations in terms of the same layer of linguistic representation tend to differ significantly in details. Their differences become major hindrances for machine learning algorithms, evaluation and re-usability of NLP tools. Differences in annotation are more substantial in ontology-oriented or task-oriented annotation, since the seemingly same types of annotation are performed based on different principles and guidelines which reflect diverse interests of groups which are involved in such annotation.

Our group is being engaged in text mining research for biology, and has performed text annotations of various types. The GENIA corpus, which consists 2000 abstracts (approx 0.5 million words) from MEDLINE, were annotated not only in terms of linguistic information such as POS, PTB-style phrase structure, feature structures of HPSG and co-references but also in terms of biological ontology such as named entities of various semantic classes and relations among them such as biological events.

Annotations of the GENIA corpus for Bio-NLP tasks have brought interesting research challenges which I would like to discuss in my talk. Those are as follows. All are somewhat related to different aspects of re-usability of annotated corpora, which due to the cost of annotation has become one of the central issues.

1. **Adaptation for Sublanguages**: Although many NLP tools are trained by corpora of the general domain such as PTB, texts which we have to deal in real applications tend to have very different linguistic characteristics. However, in most of applications, we cannot construct large annotated corpora comparable to the size of those in the general domain. Given the limited size of domain specific corpora, we need effective means by which we can re-use large annotated corpora of the general domain for constructing domain specific models.
2. **Re-use of task-oriented corpora for similar tasks**: Even annotations for the seemingly same tasks often differ in detail. Direct re-use of corpora annotated by different groups does not yield good results. More sophisticated approaches such as semi-supervised learning, transfer learning, etc. need to be deployed for effective sharing of ontology-based annotations.
3. **Linking linguistic annotation with ontology-based annotation**: The main assumption of NLP research is that linguistic structures of various layers, which are universal across different domains or tasks, can help extraction of task-oriented information from text. Till now, syntactic structures (e.g. phrase or dependency structures) have been proven useful to extract biological events. Now, we need to show how deep semantic representations such as those proposed by PropBank, FrameNet, etc. are effectively used for task-specific extraction.

[9] Download the presentation at: http://www.flarenet.eu/sites/default/files/Tsujii_Presentation.pdf

4. **Inter-annotator discrepancy**: Annotations of deeper layers tend to be more dependent on individual annotators than those of shallow layers. To reduce discrepancies among annotators, one has to construct mini-theories which link surface language forms with ontology. In fact, construction of such mini-theories seems to be equivalent to development of semantic mini-theories based on instances of annotations in text. Such evolutional development of semantic theories via annotation is much subtler than that of syntactic theories and has to be supported by annotation tools which maintain meta-descriptions attached to annotation instances.

Dialogue corpora remain a problem[10]

Yorick Wilks – University of Sheffield, UK
*www.companions-project.org

I want to raise for discussion an issue connected to projects I am currently involved in. It concerns dialogue corpora: in spite of the wealth of text corpora now easily accessible from the web, it has not proved so simple for projects working on dialogue forms and wanting to do machine learning over corpora of natural interchanges between speakers. Adam Kilgarriff once wrote a satirical piece arguing, with some reason, that Corpus linguistics had been distorted by constant (over)use of the Penn Tree Bank and the Wall Street Journal, the former being quite small by modern standards and the latter rather domain specific.

The general awareness that the web is now a corpus—or rather, that a very large corpus can easily be assembled from it-----has now changed all that, but the modelling of dialogue—with which I remain concerned*----has remained a Cinderella in corpus linguistics, still poor and under-resourced. There are of course large, freely available, transcribed and annotated corpora of natural dialogue like SWITCHBOARD, but they remain very general and progress seems only possible with domain-related dialogue corpora. The web is of course full of chatrooms whose dialogue one can retain, but as anyone knows who has experienced them, they lack much of the quality of real natural dialogue between people. This has the unfortunate effect that most dialogue projects still have to go and build their own relevant corpora, like one referenced by asterisk above. Even if they resort of Wizard-of_Oz techniques to generate data in a controlled way there are serious issues as to whether the product of that are of the same nature as person-to-person dialogues.

The problems in obtaining corpora are, as becomes clear very soon to those engaged in such projects, deeper than mere lack of availability: the issue is also WHAT is being modelled and its relation what the purpose of the analysis of such corpora will be. If the ultimate aim is to create/engineer automata that can hold dialogues then the data should, to be realistic, come not from person-to-person corpora like SWITCHBOARD, but from actual dialogues between people and computers. But such automata do not yet exist, and in a Wizard-of-Oz situation we capture the dialogue performance of a person who believes they are in that situation when they are not, and in relation to a human masquerading as a computer and whose performance is therefore not that of any actual, or currently possible, computer.

This situation does not arise at all in relation to modelling with text corpora, and shows that there are issues underlying corpus gathering and analysis that are not that complex, but there is no general agreement how to treat them. The easy way is just to be optimistic and carry on, assuming the distinction above do not really matter and the only way ahead is a robust practicality, that often extends to saying, once we have a trial automaton for dialogue let us bootstrap more data from random conversations with it no matter what its quality. I am dubious of this but not certain of the alternative beyond masses of project-specific data gathering which must always been of modest size.

[10] Download the presentation at: http://www.flarenet.eu/sites/default/files/Wilks_Presentation.pdf

Trends in Language Resources and New Work in ASR Data Labeling[11]

Gary Strong - JHU Human Language Technology Center of Excellence, USA

We see at least three trends in human language resources that will affect both the creation of resources to support research and the actual research to be done. All three trends respond to the rapidly increasing quantity and variety of human language data and requirements for research to address large data streams.

1. Moving away from hand-annotated language corpora to increased use of automatically labeled data.

   The cost in terms of both level of effort and money of hand-labeling is already large. In addition, the rapidly increasing number of genres requires a fresh look at how to generate training and evaluation data for human language research. Automated approaches look promising as one way to make more efficient the production of annotated corpora.

2. Moving away from corpus-based evaluation to increased need for evaluation against possibly non-stationary streaming data.

   In this era of “big data”, human language technology research and development is like many other areas of research in that it needs ever-larger sets of data with an eye toward a robustness that allows it to easily adapt to evolving data streams. New genres, enormous quantities of user query data, topic shifts, emergence of indigenous languages in the electronic world, and wide-spread use of informal communication have all contributed to a need for both a new type of technology that can adapt and new thinking about types of data upon which such technology can be created and evaluated.

3. Moving away from well-defined “test sets” for evaluation toward evaluation on other criteria accounting for robustness against previously unseen characteristics.

   Traditional human language technology research and development has been evaluated by the measurement of performance “scores” against well-defined test sets. New types of evaluation must be considered that determine technologies’ abilities to adapt to changes in fundamental characteristics of the data. Sometimes these characteristics are not foreseen within a time frame that allows creation of static test sets.

Indicative research is presented that shows promise in at least one of these trends.

[11] Download the presentation at: http://www.flarenet.eu/sites/default/files/Strong_Presentation.pdf

Going for a hunt? Don't forget the bullets![12]

Dan Tufiş - Romanian Academy Institute for Artificial Intelligence
tufis@racai.ro

In the quest for fast deploying of NL-based applications it seems that the concern on the major problems of language resources is loosing momentum and there is an overestimation of what machine learning can do in avoiding the highly expensive manual involvement in the process of building adequate language resources.

A well known slogan of the data intensive approaches to language processing (attributed to Bob Mercer) is "Better Data is More Data". The motivation behind this credo is that, due to natural redundancy in language, the main linguistic regularities would be revealed by statistical computing over huge amounts of raw data. While this continues to be true, it needs amendments: "Better Data is More Accurately Pre-Processed/Annotated Data". With the intentional ambiguity embedded into this new slogan, the idea is that exploiting the existing state-of-the-art linguistic pre-processing technologies (language identification, tokenization, tagging, lemmatization, chunking, dependency linking, text categorization etc.), available for most of the languages, the data sparseness threat is tremendously reduced and intelligent workflows architectures for automatic acquisition, annotation and indexing of linguistic data, with humans involved in the process, can lower the data hunger and increase the quality of the targeted linguistic services.

I think that there is a general agreement on the need to compromise between the automatic and human work in constructing the required language resources, but one can see a significant shift towards fully automatic ways in this endeavor. While this is not at all a bad idea, I think that what is wrong is to assume that the work is done once the automatic processing is finished! Whatever LRs are automatically constructed, they should be validated and, whenever needed, corrected by human experts. Therefore, besides innovative means for automatic acquisition and encoding linguistic data, the development of clever means for error spotting and computer-aided correction should be a major concern for LR community. Each type of language resource raises specific validation and correction problems and, thus, different instruments might be appropriate for these tasks. There are several initiatives to build comprehensive classification schemas for the plethora of existing LR (e.g. CLARIN) and such a typology might give a clear indication for the needed man-machine validation and correction systems. Most LRs today are concerned with the standard modern languages, but several other varieties of modern languages can be found in everyday life: slang (even dirty) language, emo languages, SMS languages, etc. Developing, on a systematic and concerted basis, multilingual resources for such language varieties would tremendously support, for instance, the development of "killer applications" for the social web.

In this respect, the recent years have seen some successful methods and methodologies to evaluate and improve the quality of the automatically built LR. Cleaning corpora is already an established task, largely automated, but usually followed by a human check-out. For instance, the largely used multilingual and sentence aligned corpus JRC-Acquis, contained several alignment errors and even foreign-language paragraphs in documents supposed to be entirely monolingual. A high precision language-identification program helped to identify and remove the spurious paragraphs and redoing the sentence alignment has significantly improved the quality of this corpus. There are interesting and successful methods for automatic extensions of wordform lexicons. Another relevant example is the detection and correction of tagging errors. The biased-tagging methodology and the

[12] Download the presentation at: http://www.flarenet.eu/sites/default/files/Tufis_Presentation.pdf

cross-tagging technique are among the most successful technologies in spotting errors and providing informed correction suggestions. Validating the interlingual alignment of the synsets in different language wordnets in the BalkaNet project took advantage of word alignment technology which was used on bilingual corpora to validate the EQ-synonymy relations among different literals and to detect valid synonyms absent from one or the other equivalence-related synsets. The cross-lingual transfer of various kinds of annotation is another very promising approach for automating the creation of valuable resources for new languages. Based on assumed similarity that can be modeled at various levels of a language system there have been successful approaches in transferring semantic, syntactic and even morpho-lexical annotations. While the semantic information transfer can be modeled for a class of more diverse languages, the syntactic and morpho-lexical information transfer is arguably more accurate for closely related languages. Yet, the human validation is indispensable for creating reliable resources based on cross-lingual transfer.

Several public resources (both monolingual and multilingual) are usually posted as a result of automatic processing only and, as such, in spite of their undisputable usefulness, they are not fully exploited at their real potential. Some examples at hand are the SEMCOR corpus which contains tens of thousands of POS tagging errors, the above mentioned first release of JRC-Acquis, the SentiWordNet and some other largely used resources. A good initiative would be to identify the mostly used (and thus considered most useful) resources and submit them to a systematic and professional curation process (as in digital libraries practices) and afterwards returning them to the research community.

Another issue I would like to touch upon is the value of collaborative development of language resources. Provided such an initiative is carefully moderated and supported by adequate development tools, the results can be really impressive. One can bring evidence from various initiatives, but I would mention here only two Romanian examples: the creation of the online version of the Dictionary of Contemporary Romanian Language (http://dexonline.ro/) the most used lexical resource on the web for Romanian and the monumental Thesaurus Dictionary of Romanian (https://consilr.info.uaic.ro/edtlr/wiki/index.php?title=Main_Page). These ongoing projects could have not reached the actual status without the involvement of hundreds of voluntary contributions. The construction of the Thesaurus Dictionary of Romanian started more than 100 years ago and will be completed this year. The majority of its 34 volumes and most of the associated sources from which the sense relevant citations were extracted (more than 2000 volumes) are being scanned and OCRed. For the correction of the dictionary volumes a large collaborative contribution was initiated, such that what experts remained to do was much easier. The status of the work, the monitoring of the corrections and the evaluation of the results are centralized and supported by sophisticated software, specially created for the task.

Several years ago, I argued that statistical or hybrid methods are more productive when using register-specific language training data and therefore combining an automatic classification system with several register-tuned statistical processing models can be not only a more accurate solution, but also an easy to extend one: as new varieties of language become of interest and training data is available, new categories for the text classifier can be created and appropriate statistical models associated with the new language data. Today, I keep the same belief and think that following this approach is a worth path towards building wide coverage language-based services.

One proof of concept for the benefits of a careful preparation of the input data for a highly complex task as machine translation has been recently produced by a small SEE-ERA.NET project aimed at creating resources for training statistical phrase translation systems for some Balkan languages. The training parallel corpus was very small (according to the SMT practices) containing about 1,2Mio words per language, but it was very accurately tokenized, lemmatized, tagged, chunk-parsed, pair-wise word aligned and validated by native speakers. It contained specialized language (EU legislative jargon) and was meant to experiment with translating documents of the same linguistic variety. A couple of month effort in preparing this high accuracy data has paid off. Using state-of-the-art SMT technologies (e.g. Moses factored decoding, reified alignment) the small scale experiments showed significantly higher quality than all public translation services we compared with (including Google Translate).

The lesson to be learnt from this short story is that one may turn a “going for big game hunting” enterprise into a feasible task even for small groups. Provided they have the right “bullets”!

Automatic Lexical Acquisition - Bridging Research and Practice[13]

Anna Korhonen - University of Cambridge Computer Laboratory and RCEAL, UK

There is a pressing need to develop comprehensive and accurate lexical resources for natural language systems dealing with real-world applications (e.g. high quality lexicons for information extraction and machine translation). Such resources are critical for enhancing the performance of systems and for improving their portability between domains. Currently, most lexical resources are developed manually by linguists. Manual work is costly, and the resulting resources require extensive labour-intensive porting to new tasks. Automatic acquisition or updating of lexical information from repositories of text (e.g. corpora, the web) is a more cost-effective approach to take. The approach is now increasingly viable given recent advances in Natural Language Processing (NLP) and machine learning technology. Yet, despite two decades of intensive research effort, hardly any acquisition technology has moved from research laboratories into widespread application. This holds back the development of language technology and makes it increasingly difficult to obtain funding for the research area. We can tackle this situation by concentrating research in viable areas which can benefit real world applications and act as a proof of concept for the line of research:

<u>Focus of lexical acquisition</u>

To date, research has been conducted in various areas of lexical acquisition, ranging from shallow to deep (e.g. terms, collocations, subcategorization frames, lexical-semantic classes, diathesis alternations, predicate-argument structures, word senses). While considerable research effort is required to improve performance in most areas of lexical acquisition, it is important to concentrate effort on the types of lexical information which can be acquired from large data sets with promising accuracy and which we know can benefit real-world applications the most.

<u>Acquisition techniques</u>

One of the biggest current research challenges is to improve the accuracy of existing techniques further and to replace small-scale techniques with more powerful and portable techniques. Without this leap, the technologies will always be limited in what they can achieve. For example,

- Instead of focusing on one type of lexical information (e.g. syntactic), we could integrate the acquisition of different types of lexical information (e.g. syntactic and semantic) so that they can support each other.

- Instead of conducting incremental research using existing methodology which we know will not transform the field, we could actively search for better suited and developed methodology in neighbouring fields where much of the existing methodology originally comes from (e.g. machine learning, engineering, physics).

- Instead of hoping that quantity equals quality, we could investigate the optimal balance between the two, and where quantity exceeds quality, develop sophisticated filtering techniques.

- Instead of developing approaches for supervised domain adaptation, we could focus on more realistic domain-adaptation which can deal with a small amount of training data. We should also investigate the minimum effort required to obtain the training data from users, the web, etc.

[13] Download the presentation at: http://www.flarenet.eu/sites/default/files/Korhonen_Presentation.pdf

Multi-lingual acquisition

Much of the currently available technology has been evaluated for major languages (or for English) only. Evaluating the applicability of the techniques to other languages would be critical for both theoretical and practical reasons; for 1) improving the accuracy, scalability and robustness of the techniques, 2) advancing work in other languages, 3) gaining a better understanding of the language-specific / cross-linguistic components of lexical information, and 4) improving the performance of (multilingual) NLP applications (e.g. MT, IE).

Evaluation and real-world application

Many techniques are evaluated against the same lexical resources which (being incomplete, inaccurate, unsuitable for domains, and lacking frequency information) have motivated the very development of lexical acquisition. There is a need to investigate how to obtain more accurate evaluation data with the aid of users, experts, and automatic methods. Also, although automatically acquired lexical (frequency) information is potentially useful for many applications, its practical usefulness remains largely undemonstrated. It would be critical to conduct evaluation in the context of real-world tasks (e.g. information extraction, machine translation, text classification) on general and domain data, and within and across languages.

Recent research shows that even when not fully accurate, automatically acquired lexical information can be useful. It is therefore important to move beyond experimental research and use the most promising of the current technology to acquire lexical resources where they are needed the most in both research and development. Equally important is to use the techniques to obtain (statistical) information for improving and tuning existing manually built lexical resources for different tasks. For maximum impact, we should make the techniques and resources developed available for wider academic and industrial communities and encourage their use e.g. via the internet.

The Democratisation of Language Resources[14]

Gregory Grefenstette - EXALEAD, S.A., France

In Kevin Costner's 1989 film "Field of Dreams" he heard a voice telling him to build a baseball diamond in his cornfield. The voice said something like "Build it and they will come".
http://us.imdb.com/title/tt0097351/synopsis and this "build it and they will come" has motivated a lot of people in the Web 2.0 world. It is the idea of building something without actually knowing why you are building it, just believing that something will happen. In a sense this is what happened with Google Earth. Google built it without having any idea of what people would ultimately do with it. And people (young internet people) never stop exploring new ways to use it.

I think we should do the same thing with language resources.
I propose that we build, for each language, a long list of all the words, and all the forms of words for the language.
This is not impossible.
The Oxford English dictionary has 301,100 entries.
The Grand Robert has 100,000 French entries.
The Duden has 130,000 German words.
If we list all the forms of each word, we will have less than a million words in each language.
Most forms can be generated automatically.
We can imagine a large table for each language, one word form per line.
In the first column, the word form, in the second column the normalized (dictionary entry) form of the word.
And in the third column, a simple part of speech: noun, verb, adjective, adverb, preposition, other.

acquiesçassions	acquiescer	V
acquiesçons	acquiescer	V
acquiesçâmes	acquiescer	V
acquiesçât	acquiescer	V
acquiesçâtes	acquiescer	V
acquirent	acquérir	V
acquis	acquérir	V
acquis	acquis	N
acquis	acquis	Adj

Already, if we had this resource for every European language, this resource as a free resource, entrepreneurs in each country would be able to exploit it to create new language tools: search tools, analysis tools, information extraction tools. This should exist for every language. For free.

And one thing more, we should add another column, with a few translations of the normalized word form, the most common translations.
The lists would then look like this:

acquiesçassions	acquiescer	V	acquiesce, consent, approve
acquiesçons	acquiescer	V	acquiesce, consent, approve

[14] Download the presentation at: http://www.flarenet.eu/sites/default/files/Grefenstette_Presentation.pdf

acquiesçâmes	acquiescer	V	acquiesce, consent, approve
acquiesçât	acquiescer	V	acquiesce, consent, approve
acquiesçâtes	acquiescer	V	acquiesce, consent, approve
acquirent	acquérir	V	acquire, purchase, buy
acquis	acquérir	V	acquire, purchase, buy
acquis	acquis	N	achievement, acquirement
acquis	acquis	Adj	acquired, learned

These translations do not have to be perfect, or complete, but if the larger internet community had this resource, we would find better and better versions of the lists, some of which could be managed in Wiki style. There is currently a wiktionary project, but it is incomplete in the number of words, and ill structured, and cannot be exploited as these simple lists can be, by a computer.

This I think is something that can be done quickly, without much public money (currently a translator is paid 10 euro cents per word, and we are talking about a few hundred thousand words).

One more column can be added automatically, the relative counts of each word form. Here are the Google counts of the words in pages also containing +la +le +que +si

3	acquiesçassions	acquiescer	V	acquiesce, consent, approve
540	acquiesçons	acquiescer	V	acquiesce, consent, approve
77	acquiesçâmes	acquiescer	V	acquiesce, consent, approve
255	acquiesçât	acquiescer	V	acquiesce, consent, approve
6	acquiesçâtes	acquiescer	V	acquiesce, consent, approve
426000	acquirent	acquérir	V	acquire, purchase, buy
5290000	acquis	acquérir	V	acquire, purchase, buy

I outlined this in a paper

> Gregory Grefenstette “Cross-Language Resource Needs for Internet Commerce” in **Trends in Special Language & Language Technology**, ed: Madeline Lutjeharms, Rita Temmerman, Antwepr, the Nethrlands, pp 177-198, 2001.
> ISBN: 9789002177415

There is no technical reason why this has not been done, only politics.

Web3.0 and Language Resources[15]

Marta Sabou - Knowledge Media Institute (KMi) -The Open University, Milton Keynes, United Kingdom
R.M.Sabou@open.ac.uk

The Web is constantly evolving to meet the requirements of its increasingly important role in our society. The first, largely textual generation of the Web has evolved into the so-called social Web, or Web2.0, where content is primarily contributed by users in the form of wiki pages (most notably Wikipedia), blogposts and various tag-annotated multimedia resources (images, URL's, video) in social sharing sites (e.g., Flickr). Technology experts predict another transformation into Web3.0. – a social Web that leverages from the benefits of large-scale, semantic-annotations provided by the Semantic Web (SW). Indeed, in parallel to the Web2.0 movement, the intensifying SW initiative is generating a large body of semantic data (ontologies and annotations), which is now available online and easily accessible through Google-like semantic Web gateways.

The importance of the Web for language technology as a large-scale, heterogeneous and up-to-date collection of data has been recognized early on. The large body of research in using Web1.0. for the benefit of LRs has already been extended with research focusing on Web2.0 data. For example, Wikipedia and folksonomy tagspaces are used to estimate relatedness between terms [1, 2]. *We believe that besides deepening research on the frontier of Web2.0 and LRs, the next important wave is in exploring Web3.0. resources.*

In our lab, we are carrying out pioneering work on exploring and combining resources specific to Web3.0., namely social and semantic web data. Specifically, we exploit SW data in the form of online ontologies to derive semantic relations between terms. This technology has been successfully applied in a variety of SW related tasks that traditionally would rely on established LRs such as WordNet: ontology matching (finding relations between the terms of two ontologies) and ontology evolution (extending a given ontology with new terms). In both tasks we obtained high precision values (above 70%), but coverage is still limited in some topic domains [3]. We have also applied this research to folksonomy tagspaces, with the aim to transform semantically weak, tag-based annotations of media resources into a semantically richer structure. We accomplish this both by relying on traditional LRs such as WordNet but also by employing SW ontologies. Specifically, our algorithms (FLOR and Scalet) associate tags with ontological entities and link them to one another through a variety of semantic relations. The result is a richer annotation for media resources, which can support more powerful forms of search and browsing. An initial evaluation of such enrichment process has yielded high precision but a relatively low recall, due primarily to the aforementioned relative sparseness of semantic web data. Initial user based evaluation of the enhanced resource search functionalities has produced positive feedback [4].

[15] Download the presentation at: http://www.flarenet.eu/sites/default/files/Sabou_Presentation.pdf

In conclusion, it is our view that the results from these initial experiments on using Web3.0 resources as novel LRs encourage further research in this direction. In particular, we need to explore how to combine these various resources with more traditional LRs in order to obtain optimal results. For our particular focus on exploiting online ontologies this could lead to increasing recall values. Additionally, we need to identify methods for exploring these novel LRs, possibly by adapting methods used for more traditional LRs (e.g., WordNet).

[1] M. Strube, S.P. Ponzetto, Simone P. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In Proc. of the 21st National Conference on Artificial Intelligence, 2006.
[2] G. Stumme , D. Benz , C. Cattuto , A. Hotho. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In Proc. of ISWC, 2008.
[3] M. Sabou, M. d'Aquin, E. Motta. Exploring the Semantic Web as Background Knowledge for Ontology Matching. Journal on Data Semantics, XI, 2008.
[4] S. Angeletou, M. Sabou, E. Motta. Improving folksonomy search with FLOR. Submitted for peer review, 2009.

Exploiting Croudsourced Language Resources for Natural Language Processing: 'Wikabularies' and the Like[16]

Iryna Gurevych - UKP Lab, Technische Universität Darmstadt, Hochschulstr. 10, D-64289 Darmstadt, Germany
gurevych@tk.informatik.tu-darmstadt.de
http://www.ukp.tu-darmstadt.de

The UKP Lab at the Technische Universität of Darmstadt has accomplished some explorative work on analyzing and accessing croudsourced (i.e. collaboratively constructed) web-based language resources, such as a set of the Wikipedia-based resources and Wiktionary which are called "Wikabularies" in this contribution. The resulting sources of background knowledge have been effectively utilized in two different Natural Language Processing tasks, such as information retrieval and computing semantic relatedness of words. Further studies involving the use of "Wikabularies" in the tasks of sentiment lexicon induction for opinion mining and paraphrase generation for enhanced Question Answering are underway. The high-performance Java-based APIs to access multilingual editions of Wikipedia and the English and German editions of Wiktionary are freely available to the research community at http://www.ukp.tu-darmstadt.de/software/.

Keywords: Wikipedia, Wiktionary, semantic relatedness of words, information retrieval.

"Wikabularies" are collaboratively constructed web-based resources that emerge as the result of collective intelligence on the basis of the Wiki technology and can be utilized as substitutes of conventional language resources in a variety of Natural Language Processing (NLP) tasks. The resource that has gained the greatest popularity in this respect so far is Wikipedia. However, other resources recently discovered in NLP, such as folksonomies, the multilingual collaboratively constructed dictionary Wiktionary, or Q&A sites like WikiAnswers or Yahoo! Answers are also very promising. While such croudsourced resources, primarily Wikipedia, have considerably influenced the NLP community when used as substitutes for conventional semantic resources, the properties of "Wikabularies" and the consequences of the collaborative approach to their construction are not yet systematically studied and well understood.

The Ubiquitous Knowledge Processing (UKP) Lab at the Technische Universität Darmstadt performed a comprehensive analysis of different parts of Wikipedia that can be used in NLP, such as the Wikipedia article collection, the Wikipedia category graph and the Wikipedia article graph, and of the user-constructed lexical semantic resource and dictionary Wiktionary. We conducted comparative studies of graph-theoretic properties of "Wikabularies" and conventional language resources such as wordnets. We developed representational interoperability mechanisms for mapping between modelling

[16] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Gurevych_Presentation.pdf

entities in these resources. Currently, we work on the content interoperability of “Wikabularies” and wordnets, which involves mutual word sense disambiguation, coverage overlap, and other aspects of complementarity. For accessing the identified types of lexical semantic information in “Wikabularies”, we implemented high-performance, Java-based Wikipedia and Wiktionary APIs that are used by numerous research groups and industrial research labs worldwide. The software is freely available to the research community from our website http://www.ukp.tu-darmstadt.de/software/ and allows for an easy integration of these language resources into NLP systems.

We utilized the “Wikabularies” in the CLEF’08 competition. When a statistical model based on Lucene was combined with our semantic models using Wikipedia and Wiktionary, the mean average precision increased by 9% for English, 15% for German, and 16% for Russian in the domain-specific monolingual retrieval task. In the bilingual retrieval task, we used cross-language links in Wikipedia, whereby the terms were directly mapped to concept vectors in the target language, and achieved an increase of mean average precision by 35%. Furthermore, we performed experiments on generalizing the Explicit Semantic Analysis method proposed for Wikipedia to be used with conventional lexical semantic resources and Wiktionary. One interesting finding is that Wiktionary was the best background resource in the word relatedness ranking task and performed comparably to other resources in the word choice task. Currently, we work on further NLP applications utilizing “Wikabularies”, such as sentiment lexicon induction for opinion mining and paraphrase generation for enhanced Question Answering.

The benefits of using web based resources come along with new challenges, such as their interoperability with existing language resources and the quality of the encoded lexical semantic and factual information. As collaboratively created resources lack editorial control, they are often incomplete and display imbalanced coverage. The quality of “Wikabularies” is questioned in many cases, and the information extraction remains a complicated task due to the incompleteness and partly irregular structure of the content. For Wiktionary, sophisticated parsers have to be designed for each language-specific Wiktionary edition as there is no uniform page structure. The above listed challenges actually present a chance for NLP techniques to improve the quality of web based semantic resources. We therefore started to work on techniques for NLP based Wiki enhancements that utilize text segmentation, keyphrase extraction and link prediction techniques to guide the "crowds" during the resource construction to be better suited for being used in NLP in return (see our Wikulu, i.e. self-organizing wiki project, at http://www.ukp.tu-darmstadt.de/projects/wikulu/).

Acknowledgments

The author thanks the UKP Lab’s staff for valuable contributions, especially Torsten Zesch and Christof Müller. This work is funded by the German Research Foundation (GU 798/1-2, GU 798/1-3, and 798/3-1), the Volkswagen-Foundation (I/82806), and the Klaus-Tschira Foundation (Project-Nr. 00.133.2008).

S3 – Evaluation and Validation

Chair: Jan Odijk - Rapporteur: Joseph Mariani

Introduction

The topic of this session encompasses two major issues:

- Evaluation and validation of the quality and quantity of Language Resources (LRs) which are produced for a given objective (conduct research investigations, develop a product, etc);
- Evaluation of Language Technology (LT) and production/distribution of the LRs which are necessary for developing and testing the corresponding LT.

LRs

Validation of a LR entails checking whether it has been created in accordance with its specification or documentation; it is an essential ingredient to assess the quality of LRs. Validation is systematically applied in programmes and projects in which it is known in advance that LRs will have to be distributed to others, but very often still neglected outside of such a context. The focus in validation has so far been on *formal* validation, and a systematic approach towards this kind of validation has been developed and applied to a range of resources. Though *content* validation has been applied in some cases, this has been tentative and somewhat ad-hoc since, differently than for *formal* validation, there is no established methodology for *content* validation.

In this session we intend to assess the situation around formal validation and inventory what new needs and trends there are in this respect, but we especially also hope to dig a little deeper into the problems that *content* validation poses, and how they can be overcome: Are there fundamental differences between formal and content validation or can they be approached in the same manner? What elements are lacking to make a systematic approach to content validation possible? How can we stimulate that such validation is made a systematic ingredient in the production of LRs? Etc..

Evaluation of LRs relative to a certain objective (conduct research investigations, develop a product, etc), is an assessment of whether the LR is suited for this objective. Have the LRs produced in the last decades indeed been used for the objectives they were intended for? And were they successful? What can we learn from this for future resources? Can useful LRs be created without a specific objective in mind? Are resources created with such an objective in mind not too limited in scope given the amount of effort and money invested in them? Are LRs with pretty wide objectives such as BNC and the Dutch Spoken Corpus useful and for which objectives are they actually being used?

LT

In the area of LT evaluation, we have witnessed several developments over the past twenty years, starting from the DARPA initiative in the mid 1980s which relied on NIST for conducting the evaluation campaigns and created the LDC for making the necessary LR available. Starting from the evaluation of Automatic Speech Recognition systems, it was generalized to many areas of spoken and written language processing, and to multimedia/multimodal data. Based on the same approach, several evaluation campaigns have been organized, e.g. CLEF (Europe) , TREC (US), ACE (US), NTCIR (Japan), Senseval/Semeval, EVALDA (France), EVALITA (Italy), N-BEST (Netherlands and Flanders), etc..

In the area of Machine Translation, automatic evaluation methods have been recently proposed with the specific evaluation data they require and their associated metrics, with several variants (BLEU, NIST, TER, ...), but the question of how to measure the quality of translation is still open and a matter of discussion (see the recent NIST campaign on evaluation of MT evaluation metrics). The same issue is present in other fields (e.g. Question Answering with metrics such us RR, Q-measure, etc.).

It is also important to consider the distribution of packages making it possible for researchers and industrials to evaluate the quality of their results after the evaluation campaign.

Discussion, Objectives, FLaReNet Claims

In this workshop we want to reflect on these developments, share (good and bad) experiences and look to the future. Are there new needs, new trends?

Questions

LRs

3.1. Which LR validation/evaluation methods are there already? Are such methods lacking for specific types of resources?

3.2. Can formal validation and content validation be approached in the same manner, or are there fundamental differences between them?

3.3. How can we measure the quality of LRs? What do we mean by quality?

3.4. Have the LRs produced in the last decades indeed been used for the objectives they were intended for? And were they successful? What can we learn from this for future resources?

3.5. Can useful LRs be created without a specific objective in mind?

3.6. Are resources created with such an objective in mind not too limited in scope given the amount of effort and money invested in them?

3.7. Are LRs with pretty wide objectives such as BNC and the Dutch Spoken Corpus useful and for which objectives are they actually being used?

LT

3.8. Are the models as used in the various evaluation campaigns the right model? Are adaptations needed? Are there good experiences we should promote, or bad experiences that should be shared so that they can be avoided by others?

3.9. Are there urgent needs for evaluation data or tools?

3.10. Are there new trends or desires in methodologies for carrying out evaluation, and if so do they require new types of evaluation resources (data, tools, metrics)?

3.11. Should the quality of the process of creating technology play a role in an overall evaluation, and if so, what recommendations can be made in this domain?

3.12. What should be the business model attached to LT evaluation? Should it be fully/partly supported by public funds? Should it be handled by public organizations? Should it be a prerequisite for any participation in public programs? How to port evaluation campaigns to other languages?

The “Standard Deviation” of LR Quality[17]

Henk van den Heuvel - SPEX/*CLST, Radboud University Nijmegen*
H.vandenHeuvel@let.ru.nl

Two approaches have been developed to assess the quality of Language Technology (LT) and Language Resources (LRs) during the last decade: Evaluation and validation. The term *evaluation* is used for the quality assessment of LT (systems and tools). The term *validation* is used for quality assessment of LRs. In that context, validation is traditionally defined as the check of a LR against its specifications (often derived from the documentation). However, as I see it, LR validation will increasingly take the shape of LT evaluation in the next decades. The crucial elements in this shift are the increased facilities for standardisation and the improved performance of LT itself. I’ll briefly explain this.

The traditional notion of LR validation was developed in the SpeechDat framework. Protocols and procedures were devised to produce LRs that were of equal quality. Uniformity had to be achieved by standardisation. What we have learnt from that experience is that standardisation is a means to warrant:

- (re)usability
- data merging / interoperability
- multi-linguality
- automatic validation

Thus, standardisation is a key to LR quality. Consequently, there is direct relation between validation and standardisation. **Standardisation will become a dominant means for validation** in the future.

The components of a LR are (apart from the data proper, such as audio-files and texts):

1. its metadata
2. its documentation
3. its contents (transcripts, annotations)

The keys to standardisation are the definition of appropriate metadata sets and automatic content generation. Existing LRs usually contain most of their *metadata* (such as speaker information, recording characteristics) in some formal structure that is machine readable and that can easily be converted and/or aggregated into a standard metadata set formalism. Standards for metadata sets are developed by such initiatives as OLAC, IMDI, ISOcat, and most recently CLARIN. Efforts should be made to standardise the pointers to fragments of data as well preferably through the use of persistent identifiers.

The *documentation* is crucial for the proper use of a LR. Therefore, standardising the documentation is one of the primary challenges for the future. As I see it, this standardisation of documentation is directly related to the metadata challenge. After all,

[17] Download the presentation at:
http://www.flarenet.eu/sites/default/files/van_den_Heuvel_Presentation.pdf

documentation is informal metadata, and metadata is formalised documentation. Thus creating standardised metadata sets is the way to standardise documentation as well.

If we have the *content* in some (standard) formalised framework, then this does not say anything about the quality of the content itself. The creation of content (annotations) is essentially a matter of human effort at present. And so is its validation. What we should aim at is to deploy language and speech technology to generate this content automatically, e.g. ASR for transcripts. The use of tools will also make it easier to adhere to annotation standards. Therefore, in the future, content validation will more and more boil down to the evaluation of the tools that created the annotations. In this way, LR validation will increasingly evolve into LT evaluation.

As a result of this shift the “standard deviation” will be become the fundamental measure of LR quality.

Towards more effective LR validation[18]

Florian Schiel, Bavarian Archive for Speech Signals (BAS), University of Munich, Schellingstr. 3, 80799 München,Germany
schiel@bas.uni-muenchen.de, www.bas.uni-muenchen.de

The BAS is performing LR evaluations internally and for other institutions on a regular basis. Therefore, the following regards only to LR validation and evaluation.

1. 'Automatic' vs. 'manual' validation
The distinction 'formal' vs. 'content' is impractical; we prefer for easier budgeting to distinguish between established methods that can be carried out automatically and methods that require human intervention. Furthermore we deem the validation of the documentation and meta data a separate task. A typical LR validation therefore consists of the following points:

- identification of reference from
  – specification
  - documentation
  - established good practise (e.g. SPEX or BAS guidelines)
- validation of documentation and meta data
- automatic validation
- manual validation
- quality assessment

What is to be done here?

- *development of more automatic tools for validation; examples are: automatic fault detection (clipping, noise,...), parsers for established annotation formats, detection of segmentation errors (using statistical methods).*
- *establish 'good practise' recommendations for manual validation*
- *minimize effort for manual validation*

2. Quality of an LR
A validation can result in an overall quality assessment, but these are not comparable across heterogeneous LRs. In practise a validation report contains a weighted summary of found problems (priorized list), which allows the producer/maintainer of the LR to take action or the prospective user of an LR to decide whether it is suitable for his/her purposes.
Keep in mind that a validation is in reference to a specification: if a specification allows wide margins for a certain error, the validation does not evaluate this specification.
What is to be done here?

- *establish 'good practise' recommendations about priorities*
- *establish 'good practise' recommendations for specifications (e.g. BAS guidelines)*

3. Evaluation of an LR
Evaluation of an LR makes only sense in reference to a specific objective.

[18] Download the presentation at: http://www.flarenet.eu/sites/default/files/Schiel_Presentation.pdf

This can be done, is difficult and expensive, but does not give any overall quality assessment for the LR, since the LR might be perfectly suited for another purpose than the original objective, which cannot be foreseen by the evaluator. In practise evaluations of a LR are hardly ever done because

- it is hard to agree on evaluation criteria
- a meaningful evaluation requires that the evaluator acts as a real 'user' of the LR

What is to be done here?

- *nothing*

For detailed guidelines regarding validation of speech corpora please refer to:
www.bas.uni-muenchen.de/forschung/BITS/TP2/Cookbook

References:

Schiel, F. (2003). *The Validation of Speech Corpora* . Bavarian Archive for Speech Signals, (ISBN: 3-8330-0700-1).

Schiel, F., Draxler, Chr. (2003). *The Production of Speech Corpora*. Bavarian Archive for Speech Signals, (ISBN: 3-8330-0700-1).

Henk van den Heuvel, L. Boves, E. Sanders: Validation of Content and Quality of Existing SLR: Overview and Methodology. ELRA/9901/VAL-1 Deliverable 1.1, Jan 2000.

Evaluation of Technology for Multilingual Information Access: the Next Step[19]

Carol Peters - ISTI-CNR, Pisa, Italy

There is general consensus that well-organised worldwide evaluation initiatives contribute significantly to the building of strong research communities, advancement in state-of-the-art, and industrial innovation in a given domain. In the Information Retrieval (IR) field, there are currently three major internationally recognized activities of this type: the Text Retrieval Conference[20] (TREC) series organized by the US National Institute of Standards and Technology (NIST); the NTCIR Evaluation of Information Access Technologies[21] organized by the National Institute of Informatics, Japan; the Cross Language Evaluation Forum[22] (CLEF), recently supported by the European Commission under FP7[23].

CLEF has been running for almost ten years now with the main goal of sustaining the growth of excellence in language processing and multilingual information access (MLIA) across language boundaries within the global context of the multilingual Web. Over the years, strongly motivated by the need to promote the study and utilisation of languages other than English on the Internet, a core network of research institutions involved with CLEF, with some support for the central coordination mainly from the DELOS Network of Excellence for Digital Libraries, has produced the following significant results:

- Creation of a very active multidisciplinary international research community, with strong interactions with both TREC and NTCIR including coordination of schedules and activities;
- Investigation of core issues in MLIA which enable effective transfer over language boundaries, including the development of multiple language processing tools (e.g. stemmers, word decompounders, part-of-speech taggers); creation of linguistic resources (e.g. multilingual dictionaries and corpora); implementation of appropriate cross-language retrieval models and algorithms for different tasks and languages;
- Creation of important reusable test collections and resources in diverse media for a large number of European languages, representative of the major European language typologies;
- Significant and quantifiable improvements in the performance of MLIA systems;

However, since CLEF began the associated technologies, services and users of multilingual IR systems have been in continual evolution, with many new factors and trends influencing the field. For example, the growth of the Internet has been exponential with respect to the number of users and languages used regularly for global information dissemination. The expectations and habits of users are constantly changing, together with

[19] Download the presentation at: http://www.flarenet.eu/sites/default/files/Peters_Presentation.pdf

[20] http://trec.nist.gov/

[21] http://research.nii.ac.jp/ntcir/

[22] http://www.clef-campaign.org/

[23] CLEF is currently run as an activity of the TrebleCLEF Coordination Action; TrebleCLEF is responsible for disseminating the results of CLEF to application and industrial communities, see http://www.treblecle.eu/

the ways in which they interact with content and services, often creating new and original ways of exploiting them. Language barriers are no longer seen as inviolable and there is a growing dissatisfaction with the technologies currently available to overcome them.

This constantly evolving scenario poses challenges to the research community which must react to these new trends and emerging needs. CLEF initially assumed a user model reflecting simple information seeking behavior: the retrieval of a list of relevant items in response to a single query that could then be used for further consultation in various languages and media types. This simple scenario of user interaction has allowed researchers to focus their attention on studying core technical issues for CLIR systems and associated components.

If we are to continue advancing the state-of-the-art in multilingual information access technologies, we now need to rethink and update this user model. We have to study and evaluate multilingual issues from a communicative perspective rather than a purely retrieval one. We need to examine the interactions between four main entities: users, their tasks, languages, and content in order to understand how these factors impact on the design and development of MLIA systems. It is not sufficient to successfully cross the language boundary, results must be retrieved in a form that is interpretable and reusable. Future cross-language system evaluation campaigns must activate new forms of experimental evaluation - laboratory and interactive – in order to foster the development of MLIA systems more adherent to the new user needs. We need a deeper understanding of the interaction between multicultural and information proactive users, multilingual content, language-dependent tasks, and the enabling technologies consisting of MLIA systems and their components.

At the same time, benchmarking efforts must prove their usefulness for industrial take-up; evaluation initiatives risk being seen as irrelevant for system developers if the data they investigate are not of realistic scale and if the use cases and scenarios tested do not appear valid.

Future editions of CLEF should thus introduce a new series of evaluation cycles which move beyond the current set-up, impacting on:

- Methodology definition: evolution of the current evaluation paradigm, developing new models and metrics to describe the needs and behavior of the new multicultural and multi-tasking users;
- System building: driving the development of MLIA systems and assessing their conformity with respect to the newly identified user needs, tasks, and models;
- Results assessment: measuring all aspects of system & component performance including response times, usability, and user satisfaction
- Community building: promoting the creation of a multidisciplinary community of researchers which goes beyond the existing CLEF community by building bridges to other relevant research domains such as the MT, information science and user studies sectors, and to application communities, such as the enterprise search, legal, patent, educational, cultural heritage and infotainment areas;
- Validation of technology: providing a reasonably comprehensive typology of use cases and usage scenarios for multilingual search, validated through user studies, to enable reuse of appropriate resources and to enable common evaluation schemes;

- Technology transfer: guaranteeing that the results obtained are demonstrated as useful for industrial deployment.

Achieving this goal will require further synergy between various research communities including machine translation, information retrieval, question answering, information extraction, and representatives from end user groups. Furthermore, if this programme is to be implemented, it is clear that CLEF – or any similar evaluation initiative – cannot operate only via a voluntary networking basis; a solid underlying management and coordination structure is crucial in order to ensure that the programme of activities is viable, consistent and coherent and that CLEF can successfully scale up and embrace new communities and technological paradigms.

The presentation will focus on:

- the importance of evaluation activities in the promotion of system development
- the achievements and limits of the Cross Language Evaluation Forum
- proposals for future directions that guarantee continuing progress in MLIA system research and development and effective transfer of the results to application communities.

Can Evaluation Be Application-Independent?"[24]

Bente Maegaard, Københavns Universitet - CST, DK

[24] Download the presentation at: http://www.flarenet.eu/sites/default/files/Maegaard_Presentation.pdf

Language Technology Evaluation: which Funding Strategy?[25]

Edouard Geoffrois, DGA

Introduction

Quantitative evaluation or measurement is at the heart of experimental sciences and is necessary to drive scientific and technological progress. In the case of Language Technologies (LT), objective evaluation requires a specific organization often refered to as evaluation campaigns. This organization has been in use for more than 20 years in the DARPA/NIST programs. It has proven to be attractive to European research teams participating in these programs and naturally started to be used in Europe. However, it has not developed to the same extent there: in practice the evaluation efforts are more limited and scattered, relying more on local if not individual initiative. As a side effect, evaluation also remains a subject of more debate. One can therefore step back and wonder what are the basic reasons to organize evaluation campaigns, what are the impediments to such an organization, and what can be done to overcome them.

The rationale for evaluation campaigns

When analysing the reasons for organizing evaluation campaign, one can distinguish two types of arguments corresponding to two questions: why evaluation is beneficial, and why should it be organized in the form of campaigns?

Arguments about the benefits of evaluation are generic ones: It allows researchers to objectively compare approaches and to reproduce experiments, and more generally to make issues explicit, to validate new ideas and to identify missing science; It is also an important tool to judge funding efficiency and to determine the maturity of the developments for a given application.

The reasons for organizing evaluation in the form of campaigns in the case of LT aremore specific to the domain, and aremanyfold. First, the results must be measured using common test data and protocols in order to be comparable. Furthermore, since developing the technologies implies some learning, the test data should not be known in advance. But for the sake of reproducibility and scientific progress, this data should also be published after the test and discussed. As a further consequence of having some data unknown before the test but discussed after, the testing period should be common to all systems under measurement. All this implies a specific organization where all activities must be synchronized. Such an organization might be perceived as complex at first sight, but results from intrinsic properties of the domain and is necessary for a sound evaluation.

The lack of LT evaluation infrastructures

LT evaluation can be analyzed, in economic terms, as a market failure. This was further analyzed in a separate article [1]. The analysis can be summarized here as follows. Since the technology is about tasks which are not yet automatized, providing the testing

[25] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Geoffrois_Presentation.pdf

infrastructure necessarily implies an important cost in human expertise. This is almost entirely a fixed cost, while the marginal cost per system under measurement is negligible. In addition, the testing infrastructure should be easily accessible to new research teams without long-term advance planning. As a result, neither the participants to the evaluation nor its organizer have a direct interest in investing in an infrastructure which will be useful to others and which does not clearly generate a future business.

In that context, the partial funding grant model, which relies on the intrinsic interest of the beneficiaries to incite them to cover the rest of the costs, does not provide enough incentive effect. In fact, the economic actors which are the most interested in funding the evaluation infrastructure are the ones who seek to foster progress in the domain as a whole. In some specific cases where there is a niche market, one actor might be dominant enough to take the lead, but for more general purpose technologies with a wide range of applications and many actors, fostering the evaluation infrastructure is the role of the research funding agencies. However, to actually do this, the adequate instruments must be available.

How to adapt the infrastructure to the needs?

The main factor to get evaluation infrastructures adapted to the research needs is to grant it 100% public funding. There are at least two different, complementary ways to do this. One is to adapt the existing funding instruments to include evaluation activities as a special case, in a similar way to program management activities. Another one is to rely on dedicated public bodies which have LT evaluation as a part of their public missions and fund their additional costs through each specific program. In any case, it is a matter of funding strategy rather than a purely technical issue.

A secondary factor to get suitable evaluation infrastructures is to tightly connect them to the research they serve, and in particular to design all types of activities together when preparing a new research program. Additionnally, since all the research activities on the same topic should share the same evaluation data and protocols, gathering similar research in single large programs can be expected to be more efficient than scattering it in different smaller ones.

Conclusion and perspectives

To summarize, objective LT evaluation is beneficial and should be organized in the form of evaluation campaigns, if possible embedded into large integrated programs. However, the traditional partial grants combined with a lack of dedicated public structures results in a shortage of evaluation infrastructures. Different funding strategies are required to ensure that these infrastructures are suited to the needs of research and development. It is therefore of critical importance to set up new funding strategies for LT evaluation in Europe to get the full benefits of the large investments in the domain.

References

[1] Edouard Geoffrois. 2008. An Economic View on Human Language Technology Evaluation, in Proc. International Language Ressources and Evaluation Conference (LREC).

Toward an Integrated Evaluation Framework[26]

Bernardo Magnini - FBK-irst, Trento, Italy

In the last years we assisted to an increasing offer of evaluation campaigns in the area of language technologies. Roughly, we can individuate two types of such campaigns: task-oriented evaluations and application-oriented evaluations.

In *Task-oriented evaluation* a single task is evaluated independently of the final application scenario. This approach tends to maximize task performance and to reuse methodologies (e.g. machine learning) through tasks. Examples of successful task-oriented evaluations are named entities recognition, semantic role labeling and word sense disambiguation. While task-oriented evaluation has generated an impressive number of initiatives, in several cases it is still difficult to understand the impact of a single component in the final scenario.

In *Application-oriented evaluation,* the overall application scenario is evaluated independently of the intermediate components involved in the process. Examples of successful application-oriented evaluations are question answering, summarization and machine translation. In application-oriented evaluation the focus case it is usually difficult to understand the role of single components. As a consequence, the approach tends to maximize global performance.

We suggest an *integrated evaluation framework* where single components are not evaluated per se, but rather for their contribution to a global application. The ideal infrastructure for integrated evaluation would be a network of web services, where each web service serves a specific component. The main expected benefits are that (i) components which contribute most will be rewarded, in term of interests, this way fostering new research directions; (ii) new metrics will be developed to evaluate single components in complex architectures; (iii) a larger amount of ablation tests both for components and for resources will be available, as well as fine grained quantitative and qualitative analysis; (iv) new methodologies for faster prototyping of final applications will be developed.

According with the above considerations, we suggest the following roadmap toward integrated evaluation (5 years):

- Develop shared communication protocols for single-task components;
- Support interoperability of single-task components within global applications;
- Set up a web infrastructure based on web services on the base of shared communication protocols;
- Promote the use of ablation tests in current and future evaluation initiatives, both for resources and for tools.

[26] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Magnini_Presentation.pdf

Evaluation: a paradigm that produces high quality language ressources

Patrick Paroubek

LIMSI-CNRS
BP 133 91403 Orsay cedex, FRANCE
{pap,anne}@limsi.fr

Abstract

We show how the paradigm of evaluation can function as language resource producer for high quality and low cost validated language resources. First the paradigm of evaluation is presented, the main points of its history are recalled, from the first deployment that took place in the USA during the DARPA/NIST evaluation campaigns, up to latest efforts in Europe. Then the principle behind the method used to produce high-quality validated language ressources at low cost from the by-products of an evaluation campaign is exposed. It finds its origins in the experiments performed after speech recognition evaluation campaigns in the USA, when the outputs of the participating systems were combined with a simple voting strategy to obtain higher performance results and also in theoretical results from machine learning, which show that one can combine low performance classifiers to obtain an improved system.

1. The Evaluation Paradigm

Apart from the ALPAC event (S. Nirenburg and Wilks, 2003), the evaluation paradigm was initially deployed in the United States in the framework of DARPA projects in 1987, with the organization of a series of evaluation campaigns for speech processing (Pallett, 2003), then for text understanding with the MUC campaigns (Hirschman, 1998) of the TIPSTER program (Harman, 1992), as well as in the scope of other programs run later by the Association for Computational Linguistics (ACL) and by the NIST.
The paradigm is based on a two step process:

- first, create textual or voice data in the form of raw corpora, tagged copora or lexicons, which are then distributed to main actors in the field of language engineering for the realization of natural applications involving natural language processing, e.g. word sense disambiguation, POS tagging, parsing, natural language database query, message understanding, automatic translation, dictation, oral dialog, character recognition, information retrieval, information extraction, question answering, opinion mining etc.
- second, the systems are tested on similar data and compared. The results of the test sessions and the discussions ensuing from the publication of the results furnish a sound basis to compare pros and cons of the various methods and systems during a workshop.

In addition to the knowledge gained about the algorithms evaluated, the close collaboration that exists between computer scientists and linguists participating in an evaluation campaign and the resulting synergy among the actors are two other benefits from the evaluation paradigm. Collaboration is required to define the common data sets and the gold standard, to propose the evaluation criteria, to define evaluation protocols, and to organize the processing of the data.
In Europe, the first event of the sort happened in 1994 in Germany with the “morpholympics” (Hauser, 1994) on morphological analyzers for German. The same year was started in France the GRACE campaign on Part-Of-Speech taggers of French (Paroubek, 2000), then there were 7 campaigns of the FRANCIL program (Mariani and Paroubek, 1999) for text and speech, the series of self-supported campaigns Senseval on lexical semantics organized by the ACL-SIGLEX working group (Edmonds and Kilgarriff, 2003), its follow-up Semeval (Agirre et al., 2007) or the more recent evaluations campaigns for Portuguese text analysis (Santos and Cardoso, 2006), as well as examples of national programs on evaluation like TECHNOLANGUE (Mapelli et al., 2004) in France with the 8 evaluation campaigns on both speech and text of the EVALDA project or the latest EVALITA (Magnini and Cappelli, 2007) in Italy with its 5 campaigns on text analysis. There were also European project which have addressed the subject of evaluation within the past few years, from EAGLES (King et al., 1996) to the CLEF evaluation series (Agosti et al., 2007).

2. ROVER

The idea to combine the output of systems participating to an evaluation campaign in order to obtain a combination with better performance than the best one is not new. To our knowledge, what now is known as the ROVER (Reduced Output Voting Error Reduction) algorithm was invented by J. Fiscus (Fiscus, 1997) in a DARPA/NIST evaluation campaign about speech recognition. He found out that by aligning the output of the participating speech transcription systems with a dynamic programming algorithm (Allison et al., 1990) and by selecting the hypothesis which was proposed by the majority of the systems, he obtained better performances than with the best system. Since, the idea gained support, first in the speech processing community (Lööf et al., 2007), where people now work on refined versions of the algorithm, using the performance of the different speech recognizers as confidence weights in the hypothesis lattice obtained by combining the different ouptuts and by applying language models to guide the final stage of best hypothesis selection (Schwenk and Gauvain, 2000). In general better results are obtained with retaining only the output of the two or three best performing systems, in which case the relative improvement can go up to 20% with respect to the best performance (Schwenk and Gauvain,

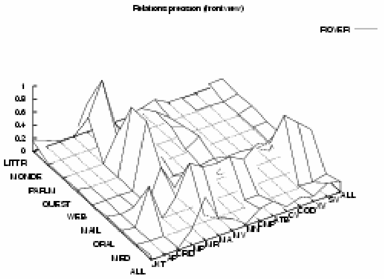


Figure 1: ROVER relative gain of performance in precision for syntactic relation annotation against the best performance

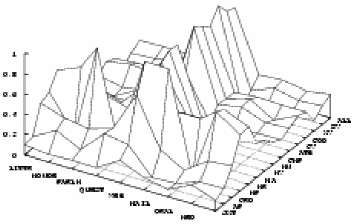


Figure 2: Difference between recall of union of all participants and best recall performance for syntactic relation annotation at EASY campaign

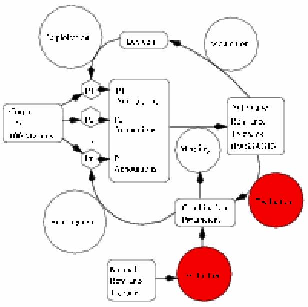


Figure 3: How PASSAGE uses the evaluation paradigm (in grey) to identify ROVER parameters in order to produce automatically a large sized treebank with high quality annotation.

2000). For text processing, examples of use of ROVER procedure are more rare, one such instance is MULTITAG (Paroubek, 2000) for POS tagging, where the algorithm was applied to provide POS tags with confidence annotation to yield a validated language resource from data produced in an evalation campaign. Machine translation evaluation is another area where ROVER algorithms are used (Matusov et al., 2006). The ROVER now begins to be tested as a ressource production procedure in the scope of the PASSAGE project where it is used to combine parses to produce linguistic information, see section 3.

3. Improving the Quality of Ressources

MULTITAG (Paroubek, 2000), a French CNRS project, had the goal of producing and making available a 1 Million words corpus annotated with POS tags out of the corpus tagged by the participants of the GRACE evaluation campaign. From the initial aligned corpus tagged by the taggers and the POS annotation mappings provided by the participants were produced the confidence measures by vote counting. Then manual validation was done first only on 38,643 forms (4%) out of the 830000 forms of the test corpus for which the system combination procedure had produced an ambiguous annotation (main morphosyntactic category or subcategory). In a second step, all the forms whose annotations contained number, gender or person information (64,061 forms of the test corpus, roughly 8%) were manually checked. Thus only less than 10% of the corpus needed to be hand checked to obtain a validated annotations.

The syntactic annotations produced by the parsers that participated to the EASY evaluation campaign gave the occasion to test a ROVER algorithm. What we found to work best was by weighting the annotation of a system proportionally to the rank the system obtained at the evaluation, in a way that the annotation of the best system could be changed only if the majority of the other systems voted against it. With this algorithm, we obtained the relative gain of performance in precision for syntactic relation annotation against the best performance shown in figure 1 (Paroubek et al., 2008). In figure 2 we show the difference between recall of union of all participants and best recall performance for syntactic relation annotation at the EASY campaign. The interesting fact about this data plot is that it is always positive, it means that the potential gain in recall by combination methods is always possible, for any kind of relation and any kind of corpus genre, provided that one can identify the right weight to give to the output of each parser. That is precisely one of the aim of PASSAGE (de la Clergerie et al., 2008) (Paroubek et al., 2009), whose aim is to use the paradigm of evaluation to identify which parameters to give to a ROVER combination procedure to produce automatically a large sized treebank following the schema given in figure 3. The theoretical grounds behind the ROVER algorithm come from Machine Learning, with Vaillant's PAC model of learning, more precisely with the work of (Javed A. Aslam and Scott E. Decatur, 1993) on boosting the accuracy of weak learning algorithms which fall whithin the Statistical Query model, a model introduced by Michael Kearns to provide a general framework for efficient PAC learning in the presence of classification noise.

4. Conclusion

We have looked at the recent history of the paradigm of evaluation both in United States and Europe and have shown that be used to produce valided hight quality linguistic annotations at a relatively low cost.

5. References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Association for Computational Linguistics, Prague, Czech Republic, June.

Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, Donna Harman, and Carol Peters, 2007. *Proceedings of the 11th Conference on Research and Advanced Technology for Digital Libraries*, chapter The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe, pages 509–512. Springer Verlag. ISBN 3540748504, 9783540748502.

L. Allison, C. S. Wallace, and C. N. Yee. 1990. When is a string like a string? In *Proceedings of International Symposium on Artificial Intelligence in Mathematics (AIM)*, Ft. Lauderdale, Florida, January.

Phil Edmonds and Adam Kilgarriff. 2003. Special issue based on senseval-2. *Journal of Natural Language Engineering*, 9(1), January.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). In *In proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–357, Santa Barbara, CA.

Donna Harman. 1992. The darpa tipster project. *ACM SIGIR Forum*, 26(2):26–28. ISSN:0163-5840.

Roland Hauser. 1994. Results of the 1. morpholympics. *LDV-FORUM*, 11(1), June. ISSN 0172-9926.

L. Hirschman. 1998. Language understanding evaluations: lessons learned from muc and atis. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 117–122, Grenade, Espagne.

Maghi King, Bente Maegaard, Jörg Schütz, Louis des Tombes, Annelise Bech, Anne Neville, Antti Arppe, Loran Balkan, Colin Brace, Harry Bunt, Lauri Carlson, Shona Douglas, Monika Höge, Steven Krauwer, Sandra Manzi, Cristina Mazzi, Ane June Sieleman, and Ragna Steenbakkers. 1996. *EAGLES Evaluation of Natural Language Processing Systems.* Center for Sprogteknologi, Cophenhaguen, october. ISBN 87-90708-00-8.

J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, , and H. Ney. 2007. The rwth 2007 tc-star evaluation system for european english and spanish. In *In proceedings of the Interspeech Conference*, pages 2145–2148.

Bernardo Magnini and Amadeo Cappelli, editors. 2007. *Evalita 2007: Evaluating Natural Language Tools for Italian*, volume IV n°2, Roma, June. Associazione Italiana Intelligenza Artificiale (AI*IA). ISSN 1724-8035.

Valérie Mapelli, Maria Nava, Sylvain Surcin, Djamel Mostefa, and Khalid Choukri. 2004. Technolangue: A permanent evaluation and information infrastructure. In *In proceedings of the 4th international Conference on Language Resources and Evaluation (LREC)*, volume 2, pages 381–384, Lisboa, Portugal, May. ELDA.

Joseph Mariani and Patrick Paroubek. 1999. Human language technologies evaluation in the european framework. In *Proc. of the DARPA Broadcast News Workshop*, pages 237–242, Herndon, VA, February. Morgan Kaufmann.

Evgeny Matusov, N. Ueffing, and Herman Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 158–165, Trento, Italy.

E. de la Clergerie, O. Hamon, D. Mostefa, C. Ayache, P. Paroubek, and A. Vilnat. 2008. Passage: from French parser evaluation to large sized treebank. In ELRA, editor, *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco, May.

Javed A. Aslam and Scott E. Decatur. 1993. General bounds on statistical query learning and pac learning with noise via hypothesis boosting. In *Proceedings of the 34th Symposium on Foundations of Computer Science*, Foundations of Computer Science, pages 282–291. IEEE, November.

David Pallett. 2003. A look at nist's benchmark asr tests: past, present, and future. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 483–488, Virgin Islands, USA, November. IEE. ISBN:0-7803-7980-2 / DOI:10.1109/ASRU.2003.1318488.

P. Paroubek, I. Robba, A. Vilnat, and C. Ayache. 2008. EASY, evaluation of parsers of French: what are the results? In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco.

Patrick Paroubek, Eric de la Clergerie, Sylvain Loiseau, Anne Vilnat, and Gil Francopoulo. 2009. The passage syntactic representation. In *In proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, January. Netherlands Graduate School of Linguistics.

Patrick Paroubek. 2000. Language resources as by-product of evaluation: the multitag example. In *In proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, volume 1, pages 151–154.

H. Sommers S. Nirenburg and Y. Wilks, editors, 2003. *Readings in Macine Translation*, pages 131–135. MIT Press, Cambridge, Massachusset. ISBN-10: 0-262-14074-8, ISBN-13: 978-0-262-14074-4, http://www.hutchinsweb.me.uk/ALPAC-1996.pdf.

Diana Santos and Nuno Cardoso, 2006. *A Golden Resource for Named Entity Recognition in Portuguese*, pages 69–79. Springer, Berlin / Heidelberg. ISBN 978-3-540-34045-4, DOI 10.1007/11751984_8.

Holger Schwenk and Jean-Luc Gauvain. 2000. Improved rover using language model information. In *In proceedings of the ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, pages 47–52, Paris, September.

Proposal to launch a Support Centre for Remote Evaluation and Development of Language Technologies[27]

Harald Höge, SVOX Deutschland GmbH

Due to the rapid increasing capabilities of the worldwide web the potential of using this web for development and evaluation of speech processing systems has to be investigated.

Currently to build large speech processing systems all needed software is developed and collected locally by an institution. Further, evaluation is done by loading down from the web test data provided by an evaluation center, which is processed locally leading to result data, which are sent back to the evaluation center.

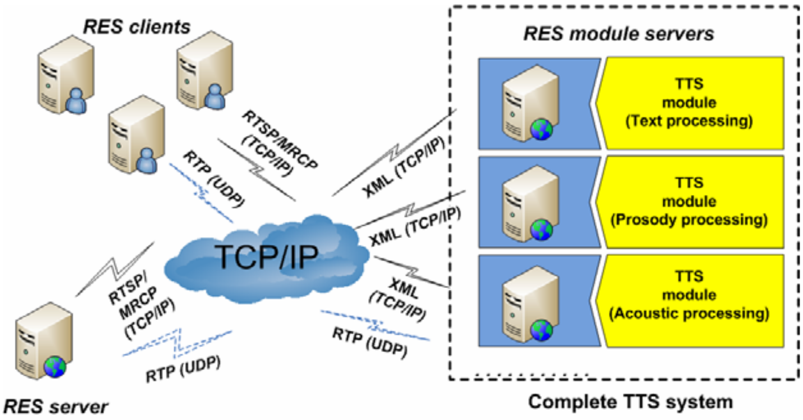
These procedures can be made much more efficient using the web. I propose that researchers and developers are supported by a web based framework which allows using and evaluating speech processing modules and systems remotely. For setting up such a framework I propose to launch a Support Centre for remote development and evaluation of Language Technologies.

A first step in this direction has been done by the ECESS[28] consortium (European Center of Excellence in Speech Synthesis). ECESS aims to speed up progress in speech synthesis technology by providing an appropriate evaluation and development framework. The key element of this framework is based on the partition of text-to-speech synthesis system into modules, the ECESS TTS modules accessible via the web. Currently a text processing, prosody generation, and an acoustic synthesis module have been specified. Such a split into modules has the advantage that the developers of an institution active in ECESS can concentrate its efforts on a single module and test its performance in a complete system using the missing modules from developers of other institutions. In this way, high performance systems can be built using high performance modules from different institutions. An evaluation methodology has been developed to assess the performances of the modules. This methodology is based on the common use of module specific evaluation criteria and module specific language resources needed for training and testing the modules. In order to evaluate the modules and to connect modules efficiently, a remote evaluation platform – the Remote Evaluation System Architecture (RESA) based on the public available software to use the web – has been developed within ECESS[29]. The RESA is based on client-server architecture as shown below.

[27] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Hoege_Presentation.pdf

[28] www.ecess.eu

[29] the RES system has been developed by Matej Roc from University of Maribor, Slovenia



Basic elements of the RESA

It consists of RES module servers, which encapsulate the modules of the researchers, a RES client, which sends data to and receives data from the RES module servers, and a RES server, which connects the RES module servers and organizes the flow of information. RESA can be used by developers to select a RES module from the web, which contains a missing ECESS TTS module needed to test and improve the performances of their own modules. Finally the RESA allows for the evaluation of ECESS TTS modules running at different institutions worldwide. When using the RES client, the institution performing evaluation is able to set-up and performs various evaluation tasks by sending test data via the RES client and receiving results from the RES module servers. Currently ELDA is responsible for setting-up evaluation using the RES client.

Specific attendance has been given to the design of RESA to integrate easily the speech processing modules of the researchers into the RES module server shell and to use easily the RES client for evaluation and testing. Currently several ECESS partners are testing the RES and ELDA is in the process to evaluate the text processing modules and acoustic synthesis modules.

Due to the design of RESA its use is not restricted to speech synthesis modules. RESA can be extended easily to other speech processing modules.

In the context of a ‘multilingual digital Europe’ I propose to create a “Centre for Remote Evaluation and Development of Language Technologies” which configures and maintains the RESA to the need of the community for evaluation and development.

Evaluation of HLT-tools for less spoken languages[30]

Cristina Vertan - University of Hamburg, Germany
vertan@informatik.uni-hamburg.de

Europe offers an unique context, not only due to the variety of spoken languages but also due to the necessity to offer a huge amount of information at least in all official U-languages.

During last years the number of mono- and multilingual systems dealing with other languages than the most frequent used ones (EN, FR, DE) increased. However most part of these systems are either developed for a particular scenario or tuned on specific (available) corpora. Consequently it is very difficult to evaluate such systems, as there are no (or less) reference test data and no (or very few) reference evaluation scores.

The evaluation of systems dealing with languages not so frequently spoken is a real challenge, which has to be urgently considered:

1. to ensure the development of qualitative similar systems across languages
2. to encourage researchers to develop tools for other languages that the frequently spoken ones.

Due to the big amount of languages, and their possible combinations one cannot built reference systems for all types of NLP tasks and all language combinations.

On the other hand following actions are a viable solution for pushing forwards the evaluation and thus development of systems dealing with less spoken languages:

- ensure a reference parallel corpora
- build reference test suites
- define reference measures for classical NLP Applications (MT, IR,CLIR, QA, etc.) – underway.
- develop when possible language independent system models (like Moses in MT)

For the moment the only existent corpus covering 22 languages in the EU is the JRC-Acquis Multilingual Parallel Corpus (http://langtech.jrc.it/JRC-Acquis.html). The main problem with this corpus is the special type of language it covers, namely law. Systems trained on this corpus will give less convincing results on other type of texts. The other way around, systems trained on different texts will deliver suboptimal results on JRC-Acquis. Similar problems can be observed on other corpora as the OPUS Corpus (http://urd.let.rug.nl/tiedeman/OPUS/). Some other corpora developed in frame of earlier EU-Projects like MULTEXT-EAST (http://nl.ijs.si/ME/CD/docs/mte-corp.html) or n-ouse projects like ROGER (Romanian - German - English - Russian corpus, http://nats-www.informatik.uni-hamburg.de) are either encoded in old fashioned standards, or unknown to the large research community, and themselves too small to be used for large-scale applications

[30] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Vertan_Presentation.pdf

We propose the building of a parallel corps for all European languages, covering different domains: law, medicine, news, turism. This should be a test-bed for any monolingual and crosslingual application. Steps for building such corpus are:

1. Unify annotation for existent parallel corpora: JRC-Acquis, OPUS, Europarl, MULTEX., MULTEX-EAST
2. Investigate all EU projects with participation from various language communities Networks of Excellence. Especially in non-technical domains project reports are translated in all participant languages (e.g. CALIMERA http://www.calimera.org/default.aspx, I*Teach http://i-teach.fmi.uni-sofia.bg/)
3. Investigate existent translations of the official EU Web sites as at least part of the information is made available in all languages
4. Collect parallel corpora produced in different projects

In a second steps reference measures have to be defines for each type of application. These measures should consider the data-set profiling aspect (see DeRoeck&Al., Invited Talk at RANLP 2005) as languages do not have the same distribution. Experience in CLEF Competitions should be a good starting point.

Special attention requires Machine Translation. The recent development of the Moses open source system, offers the possibility to develop baseline Systems for any language pair, under the assumption that a significant amount of training data (parallel corpora exist).

In conclusion a possible roadmap for bringing forwards the evaluation of systems working with less spoken languages in Europe relies on following phases:

1. Define a standard for minimal parallel corpora annotation (to be used as input data)
2. Development of a reference parallel corpus in all (most part) of the accepted EU languages.
3. Releasing of test suites (parts of this corpus) to be used in test scenarios.
4. Define standard test scenarios for evaluating (LT-systems) on these test-suites.
5. Publish A dataset profiling study, considering influence of language distribution on the statistical mechanisms for all official EU-languages and major HLT applications
6. Implementation of baseline MT-systems using Moses and testing on the reference corpus

S4 – Interoperability and Standards

Chair: James Pustejovsky - Rapporteur: Nancy Ide

Introduction

Interoperability is probably one of the most important ingredients in the glue that allows integration, sharing, interchange and reuse of Language Resources and Technologies. Interoperability of resources, tools, and frameworks has recently been recognized as the most pressing need for language processing research. Interoperability is especially critical at this time because of the widely recognized need to create and merge information and annotations at different linguistic levels. Unfortunately, there are some problems in using standards: standards are often disregarded, ignored, or considered too time-consuming and effort-demanding. Recognition of the urgency for interoperability of language resources and tools is becoming more and more critical because the multilingual scenario in Europe and new emerging data and tools for strategic languages as well as minority languages needs to be faced. We will start addressing the wider issue of interoperability and standards, as a first step towards a full LR sharing, accessibility, availability (which is a core question of the US "sister" project INTEROP-SILT).

A specific area of standardization, i.e. resource documentation will be touched. Proper resource documentation is, indeed, a pre-requisite for reuse of LRs. The contents, degree of detail, and the structure of documentation of LRs currently varies wildly. There are currently no explicit guidelines for the documentation of LRs. Some projects in which similar resources were created for multiple languages (e.g. the projects in the SpeechDat family) have used a kind of template for the resource documentation. This has many beneficial effects: the information contained in the documentation for the different resources is the same; the degree of detail is similar; it is easy to find relevant information since the structure of the documentation for the different resources is identical, etc. But the documentation, even if uniformly structured as for the SpeechDat resources, is still just text (sometimes with semi-formalized parts such as lists and tables), suitable for humans but not for software programs. Another aspect of the problem is the issue of *metadata*. Metadata for a resource are a set of formalized data that describe properties of the resource, and can therefore be seen as a formalized part of the resource documentation. Indeed, a lot of information that is usually put in the documentation should also occur in the metadata. Having the information in the metadata is crucial for searching and browsing facilities, but especially for tools and services to be able to apply on the resource.

Discussion, Objectives, FLaReNet Claims

Interoperability and Standards

In the current landscape, many different standards are already around, most of them *de-facto*, others ratified by official standardization organizations, e.g. industry standards used for localization (XLIFF), translation memories (TMX), terminology (TMF, TBX).

The recent flurry of activities within the community aimed at defining standards for LRs and LTs is the apparent reply to the recognized urgency of interoperability of language resources and tools: current initiatives in the ISO TC37 SC4 working groups, such as the Linguistic Annotation Framework, the Syntactic Annotation Framework, the Semantic Annotation Framework, the Lexical Mark-up Framework absolutely go in this direction.

Questions

4.1. What's needed now that the issue of interoperability and standardisation is particularly acute in the multilingual scenario?
4.2. How to make standard adoption more appealing? And easier?
4.3. How to show players the real advantages of standard adoption?
4.4. Are the global efforts to create linked monolingual resources (wordnets and framenets) the right way to proceed?
4.5. How to extend these success stories to other types of resources?
4.6. How to make this interlinking operational in view of multilingual applications?

Standardized resource documentation

Using a documentation template and maximizing the information in the metadata will lead to many advantages:

- Less work for the developer/documentation creator;
- Less work for the user of the resource / documentation reader;
- Easier re-use of the resource by humans;
- Easier application of tools and services on the resource in a technical language resource infrastructure or in a stand-alone configuration;
- Better, more complete and more consistent documentation;
- Higher quality resources;
- Easier and faster validation of resources.

The above claims may be true, but are not likely to work because the resource producers are often not keen to use the documentation template, because:

- They forget about the documentation template and make the documentation in their own way;
- The same as before, but at the last moment they realize about the document template, but then they do not have the time or the willingness to restructure their documentation;
- They think they can make better documentation than with the template, or they may think (right or wrong) that their important and innovative work cannot be properly described by documentation according to a fixed pre-defined template;
- They already documented the resource on their web site, and do not want to convert this to a document to be included in the resource itself.

Questions

4.7. Is it feasible to create a commonly agreed upon documentation template?
4.8. What elements should such a documentation template contain?
4.9. Are different templates needed for different resource types? If so, for which ones and why?

4.10. How can we devise such templates, test them in practice and stimulate the use among resource creators?
4.11. All parts of the documentation that should also be in the metadata should be kept out of the documentation and only be put in the metadata scheme.
4.12. All other formalizable documentation that does not fit in the metadata scheme should be represented in a formal manner, and kept out of the documentation (though the documentation will contain references to them). Possible examples are lists of possible values for attributes in the resource; list of possible tags and their interpretation, etc.). These should be stored in processable files (e.g. plain text files but not PDF) in precisely specified locations.

SILT: Towards Sustainable Interoperability for Language Technology[31]

Nancy Ide (Vassar College - DCS, USA) & James Pustejovsky (Brandeis University - DCS, USA)

Our position paper outlines U.S. participation in a network to work toward achieving interoperability among language resources (in conjunction with the EU *e*Content*plus* Programme). The resulting international effort (hereafter called "the Network") will involve members of the language processing community and others working in related areas to build consensus regarding the sharing of data and technologies for language resources and applications, to work towards interoperability of existing data, and to promote standards for annotation and resource building. In addition to broad-based US and European participation, we are seeking the participation of colleagues in Asia.

The resources and technologies to be addressed include annotated corpora (texts, audio), lexicons, ontologies, automatic speech recognizers, lemmatizers, taggers for all levels of linguistic phenomena, named entity recognizers, information extractors, etc., as well as systems for search, access, and annotation of language resources. The creation and use of these resources span several related but relatively isolated disciplines, including NLP, information retrieval, machine translation, speech, and the semantic web. The goal is to turn existing, fragmented technology and resources developed within these groups in relative isolation into accessible, stable, and interoperable resources that can be readily reused across several fields.

The major activities of the Network will be:

- To carefully survey the field to identify the resources, tools, and frameworks in order to examine what
exists and what needs to be developed, and to identify those areas for which interoperability would have
the broadest impact in advancing research and development and significant applications dependent on them;
- To identify the major efforts on standards development and interoperable system design together with existing and developing technologies, and examining ways to leverage their results to define an interoperablity infrastructure for both tools and data;
- To analyze innovative methods and techniques for the creation and maintenance of language resources in order to reduce the high costs, increase productivity, and enable rapid development of resources for languages that currently lack them;
- To implement proposed annotation standards and best practices in corpora currently under development (e.g., American National Corpus, TimeBank) to evaluate their viability and feed into the process of further standards development,

[31] Download the presentation at: http://www.flarenet.eu/sites/default/files/Ide-Pustejovsky_Presentation.pdf

testing, and use of interoperability frameworks (e.g. UIMA), and implementation of processing modules, and distributing all software, data, and annotations;

- To ensure the broadest possible community engagement in the development of consensus and agreement on strategies, priorities, and best approaches for achieving broad interoperability, through sessions, open meetings, and workshops at major conferences, together with active maintenance of and involvement in open web forums and Wikis;
- To provide the technical expertise necessary to turn consensus and agreement into robust interoperability frameworks along with the appropriate tools and resources for their broad use and implementation by means of tutorials and training workshops.

Enhanced interoperability of language resources and tools has the potential to enable a major leap in the productivity of NLP research and, consequently, language processing capabilities that can ultimately impact the use and interaction with computers. Short-term benefits include the ability to combine annotations produced by different groups to study interactions among linguistic levels, creation of lexical/semantic/ontological resources that include information relevant to different sub-domains, substantially increased access to resources and tools by members of the entire community, and rapid development of resources for languages for which NLP capabilities are only beginning or have yet to be developed. In the long term, interoperability can lead to the creation of a web-based “resource grid”, in which lexical, semantic, and ontological resources are interlinked and may in turn serve as the reference for annotations of language data. We envision the eventual creation of a massive, distributed, interlinked network of linguistic data and information, together with web services to accomplish linguistic processing “on the fly”, providing unprecedented capabilities for research and language processing.

Interoperability, Standards and Open Advancement[32]

Eric Nyberg - Carnegie Mellon University, USA

Recent efforts to promote standards for the common representation of texts, annotations and related resources (annotated corpora, lexical and semantic knowledge bases, etc.) represent an important step forward in the creation of more effective language technologies applications at lower cost. This paper argues that resource interoperability is only a first step in a general trend toward greater sharing and reuse of components in language technologies applications. When an application requires the integration of many resources and algorithmic components, it is difficult to determine the adaptability or generality of those components when the system has only been tested on a particular LT problem. When we lack the means to leverage components or resources from distributed groups, development occurs in single organizations, with significantly redundant effort.

What can help is a focus on *open advancement*: the use of shared system models, open-source components, collections of challenge problems and common evaluation metrics, so that the contribution of each technology to end to end performance can be accurately measured and the community as a whole can uniformly advance system performance on an ever broadening range of problems. Our objective is to combine formal metrics and rigorous evaluation with a collaborative research process that allows a particular research community to achieve monotonically increasing performance, while managing overall research and development cost effectively. Our recent experience with open advancement has been within the field of question answering[33], but the same principles apply to all areas of language technologies research:

- **Vision**. What are the potential benefits and related costs of open advancement for each stakeholder (industry, academia, government, etc.)?

- **Challenge Problems**. Open advancement implies a set of challenge problems that will drive improvements in the state of the art across measurable dimensions; the problems are diverse by design, so that performance gains demonstrate general advancement.

- **Open Collaboration**. We must consider a collaboration model that supports a consortium of researchers from industry, academia and government working together towards open advancement, including a shared software development process.

[32] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Nyberg_Presentation.pdf

[33] Ferrucci, D., E. Nyberg, J. Allan, K. Barker, E. Brown, J. Chu-Carroll, A. Ciccolo, P. Duboue, J. Fan, D. Gondek, E. Hovy, B. Katz, A. Lally, M. McCord, P. Morarescu, B. Murdock, B. Porter, J. Prager, C. Welty and W. Zadrozny, “Towards the Open Advancement of Question Answering Systems”, IBM Technical Report (forthcoming).

The open advancement approach is based on an iterative, collaborative research process with the following steps:

- **Establish a Shared Logical Architecture**. The collaborators share a common set of data object definitions and modular interfaces, which will be used in the construction of both individual resources, text processing modules (which implement the module interfaces by consuming/producing standard data objects) and end-to-end data flows. The collaborators also share a common framework for specifying and representing a problem (corpus, test data, answer keys, etc.).

- **Define Formal Metrics**. The collaborators share a common set of metrics which will be used to measure the performance of individual resources, modules and end-to-end data flows.

- **Define Challenge Problems**. The collaborators share a set of challenge problems using the common framework; a problem definition includes a description indicating how this problem is intended to drive innovation along particular dimensions (e.g. metrics and measurements).

- **Design Experiments**. The collaborators share a common process for open advancement, which specifies the steps to be taken in configuring an experiment, conducting an experiment, gathering measurements, and reporting system performance based on those measurements.

- **Manage Development**. The collaborators follow the common process in advancing the state of the art on the selected challenge problem(s), while continuing to refine the processing modules, end-to-end dataflow(s), and the common development process itself continuously over time. Effort is invested in those component technologies and data flows which provide the best improvements on performance for selected problems.

Open advancement will make it possible for stakeholders to answer important questions from prospective users of language technology, for example:

- What is the best data flow for my problem domain? Using which components?

- What’s the expected upper bound on performance in my problem domain?

- How can performance improve if we invest in improving the technology?

- What’s the overall cost of adapting and tuning current technology for my problem domain?

- What are the most important component technologies we should invest in, given current performance levels vs. targets?

Is the LRT Field Mature Enough for Standards?

Peter Wittenburg - Max Planck Institute for Psycholinguistics, The Netherlands

At LREC 2008 Henry Thompson pointed out that there are no agreed descriptive systems in the LRT field with the exceptions of XML and UNICODE and that we still follow the principle that what colleagues did must be wrong. With respect to the latter we may not ignore that we have to deal with all the 6500 languages exhibiting a variety of fundamentally differing linguistic properties. It is this variety that makes it so difficult to come to a unified linguistic theory. It is obvious that we cannot ignore this state description when talking about standards, linguistic theorization is still controversial and therefore an area of high dynamics. While LRT industry needs to minimize costs where possible and partly tends to adopt very simple solutions - just look at the ISO 639-1/2 code standards only covering about 500 languages - researchers, however, need to be able to deal with the given variety. Therefore we cannot ignore the difference between industry and the research domain when talking about standardization. When we include the research world in our discussions we cannot claim to be a mature field. Does it then make sense to define standards knowing that the definition process costs much time? Standards not chosen at the right level of abstraction could be outdated very quickly and could hamper scientific progress; standards at a too high level of abstraction can be useless.

The primary focus of researchers is on finding and analyzing new phenomena and along this process they don't like any form of hurdle. Saying this we can observe an antagonism for data driven research where large (virtual) collections need to be created, obviously with components resulting from different researchers, projects or institutions. On the one hand current researchers don't like to deal with standards since it is time consuming; on the other hand researchers need to adhere to standards when they want to make their resources interoperable. Currently, in the LRT field except for people that work with processing chains in NLP or ASR, most researchers are still far away from creating virtual collections, i.e. areas were new standards would be required in addition to the standards for bibliography for example.

Two essential areas need to be tackled to achieve interoperability: syntactic and semantics encoding differences. With respect to structural incompatibility we could study the variety of lexica which we are faced with, the solutions we have found so far and the benefits we can identify. In our institute we have seen that every researcher comes up with his/her own lexical structure (and of course terminology which will be touched below). Different tools are being used such as Shoebox, CHAT, relational databases, Word, Excel, some XML, etc. to manipulate such lexica and in addition different structures can be found - often not even explicitly specified. The Lexical Markup Framework that is in the ISO standardization process allows us to construct lexical structures as if we would play with LEGO bricks. We have a metamodel that allows us to represent almost any lexical structure by a simple extension mechanism. This mechanism should be mature, since some of us have analyzed lexical structures now for decades, i.e. there is a deep understanding of structures that can occur and based on this knowledge it seems that we defined a model that is flexible enough. The picture is not that clear if we ask the question who will benefit from this standardization. Of course the tool builders since they just have to build one tool and only have to implement visualization and processing components once. Also if LMF is seen as a pivot model, we can reduce the costs for conversions considerably.

What about the researchers who are the real target group? Their main interest is in simple and easy to use tools and simple is the one they are used to in our world of increasing complexity of user interfaces and features. Therefore they like to use Word and Excel; badly enough they like to use these tools also since they don't enforce them to adhere to too restrictive constraints. It's the future researchers, who may become interested in merging, filtering, searching across several lexica and who will therefore have a preference for standards based tools. As tool-builder we know that we need to be ahead of the wishes of our researchers, but we cannot expect credits etc as long as we cannot demonstrate how easy all these new tools are and as we are not able to hide complexity and restrictions due to standards. So maturity of a topic such as "lexical structure" is not sufficient, we need to produce efficient and simple tools and frameworks.

With respect to the standardization of semantic aspects we just took another step in stabilizing the model underlying the ISO Data Category Registry which specifies a flat list of concepts from our field. Although we will have a hard time to agree on the definitions and to indeed fill in all relevant concepts in all required detail, we already know now that the model actually is limited. Researchers want to include constraints dependent on the application and the context and relations. Wisely enough relations were excluded to be inserted into the DCR since otherwise one can seriously doubt whether the ISOcat task would have been feasible. We just started giving the DCR a place and except for some little islands as in the terminology domain there is no experience yet in the field. Our hope is that with the DCR's help we can shift the basis of interoperability from fixed schemas with restricted element sets placed in well-defined contexts to a registry of concepts allowing everyone to put these concepts into a wide variety of contexts. Despite from the fact that all this will only be accepted if we offer the best annotation and lexicon tools out there, we do not have any idea about the complexity of the task ahead of us. In the domain of metadata in CLARIN we just decided to go for detailed semantics of the data categories, i.e. have data categories such as "date of birth", "date of creation", etc. The price of having more data categories in the DCR is balanced with the fact that their interpretation is less dependent on the context in which they will appear. A separate relation registry mechanism will allow users to state that all these different categories have "date" as broader concept, thus allowing users to search for all kinds of dates. We don't know yet whether this principle of highly granular data category semantics will be useful for other thematic profiles in the DCR.

Given this uncertainty and the fact that in many profiles such as semantic annotation the category systems are heavily debated, the use of data categories cannot be seen as a mature topic. Nevertheless, I am convinced that we are on the right way by separating semantics and structure. However, when we now start defining data categories we need to argue from the needs of the community and not from theoretical considerations only. There is a risk that we will fail if we will not be sensitive enough. We may also fail if we integrate concepts that are heavily debated, since the acceptance of the DCR will depend on its image in the community.

Interoperability via Transforms[34]

Edward Loper – Brandeis University, USA

As the number of language resources, tools, and frameworks that are generated by the NLP community increases, it is increasingly important to ensure that these language resources and technologies are interoperable. Interoperability reduces the need for redundant development work, ensures that the best available tools can be used with any data, and makes it easier to combine different data sources. Data formatting standards are one of the most basic tools for promoting interoperability, ensuring that data and tools that have been developed independently can nevertheless be used together. A number standard data formats have been developed and introduced to the field, including IBM's UIMA (CAS), Annotation Graphs, LAF, and GRAF.

Unfortunately, there's a chicken-and-egg problem in establishing new standards. In particular, when a new standard format is proposed, there typically are not many tools that support that standard; but without such tools, there is very little incentive for members of the community to adopt the standard. On the other hand, until the new standard is widely adopted, there will not be much interest in developing tools that work with that standard.

A partial solution to this dilemma is to invest resources in making high-quality tools that support the new standard. Unfortunately, developing high-quality tools takes a lot of time and effort. A somewhat more resource-friendly approach would be to extend existing high-quality tools to support the new standard. But either solution is only partial, because a large portion of the language processing tools that are used by the community are small locally-developed tools.

A second, complementary, approach is to advocate the use of the standard data formats for interchange, allowing different resources and tools that were designed independently to be used together. In order to enable this use case, it will be necessary to develop file format transforms, which convert between existing file formats (including both officially ratified and de-facto) and the standardized file formats. Developing tools to perform these file format transforms will be significantly easier than developing or adapting language processing tools; and will allow researchers using small locally-developed tools to integrate their work with other researchers' data and tools.

One of the main goals for using the standard data format as an interchange format, from the perspective of the data format developers, is to increase awareness and use of the standards. By acting as a "nexus" for converting between existing formats, the standard formats will be used by an increasingly large portion of the community. This will pave the way for more wide-spread adoption of the standards, and hopefully the new standards will become widely adopted by the community. At that point, the use of the non-standard formats should be actively discouraged. The file format transforms will still be useful, but mainly in the role of legacy support.

[34] Download the presentation at: http://www.flarenet.eu/sites/default/files/Loper_Presentation.pdf

It is important to note that transforming linguistic data between different formats is not always straightforward. Many language resources depend on specific linguistic or theoretical assumptions, and often the chosen format depends on those assumptions. This can make it difficult to transform between different representations, because of conflicts between the assumptions used by those representations. As a simple example of this type of conflict, a data format that uses a single character to mark the part of speech of a word implicitly assumes that the number of part of speech tags is fewer than the number of characters. More subtle examples of this type of conflict involve differences in assumptions about word-level tokenization. For example, it is unclear what to do when transforming a parse tree from a format that assumes that hyphenated words are a single token to a format that assumes that they are two separate tokens. As a result of this type of mismatch, it is sometimes impossible to define exact lossless transforms between two existing formats. However, when developing the standard format, we should attempt to at least ensure that transformations from any given existing format to the standard format are lossless; and that round-trip transforms from an existing format to the standard format and back are exact.

By focusing on the use of standard data formats as a means for interchange between existing formats, we can increase the utility of the new formats, by allowing them to be used with existing data and tools; and we can increase the visibility of the standards, thereby increasing the chances that the format will be widely adopted by the community.

Ontology of Language Resource and Tools for Goal-oriented Functional interoperability[35]

Key-Sun Choi - KAIST, Computer Science Department
kschoi@kaist.edu

List of Questions

1. Language resources and tools should be usable in some goal. If each language resource and tool has specific functions, is it possible to make ontology to classify the functions and to match functions with the goals? Even can we make ontology for LR and LT's usable goals?
2. During the standardization process, is it possible to have a field-testing to put their usability in some domain and in multi-language environment? Then standardization adoption could be improved and the standards are also more practical.
3. Metadata and documentation is still in the static view of their use. Operational standards could be implemented in a manner of OpenAPI, for example. Do you think that OpenAPI is a right way to enhance the standardization and interoperability?
4. How can we make a semantic equivalence among multi-language environment? Is it possible to make a standard for that? Is it possible to operational standards to do that?
5. Is it possible to have a standard for error type of LR and LT? The developers and users can make improve a specific LR and LT to reduce the possible and discovered LR and LT.

[35] Download the presentation at: http://www.flarenet.eu/sites/default/files/Choi_Presentation.pdf

Interoperability of Language Resources and Technologies (LRT) with extra-linguistic Resources and Technologies[36]

Thierry Declerck – DFKI, Germany

While the main topic of S4-Interoperability and Standards is about how standards for language/linguistic resources, tools, and frameworks can support interoperability between those, we would like to address a further challenge: how can LRT resources interoperate with other types of (extra-linguistic) resources, towards cross-media and multimodal applications.

We are perfectly aware that the standardisation of language resources and technologies is far from being solved and constitutes as such a challenge on its own; but it could nevertheless be wise to investigate as well how language resources, tools and frameworks can interoperate with other types of resources and technologies. It might be that an in-depth investigation of the possible interoperability across domains of applications in which langue resources are playing a complementary role, can help in getting a better understanding on how one can ensure interoperability within the domain of language resources and technologies itself. For this we would try to summarize the needs that can be addressed to LRT by other kind of data processing (Multimedia Analysis, Semantic Web annotation strategy, etc) in order to improve their own performances. It seems to be a reasonable assumption that the way LRT has to deliver "data" (in fact Linguistic Knowledge) to other domains for the purpose of integration can help in describing the way components of LRT have to interact between themselves. This might be valid as well the other way round: what can LT request from extra-linguistic data in order to improve its own performance.

In this short talk we present first an overview of the actual state of standardisation of linguistic annotation, within the ISO Framework, and then switch to a brief presentation of the perception on the way language annotation can contribute to Multimedia and Semantic Web applications.

Ideally the Multimedia and the Semantic Web communities (and processes) should not be aware of the internal structure of linguistic annotations, but only of the relevant result of language processing for their purpose. So for example the annotation of an image or a video sequence with language resource can not consist in the display of complex constituency and dependency information, but on more abstract and generally understood properties, like "Who", "What", "How" etc., as this is foreseen for example in the MPEG-/ standards for Multimedia Annotation. So the problem w face here is in establishing interfaces between linguistic and extra-linguistic descriptors information structures sued in the different types of annotation. We assume that this is interfacing is probably best described in the context of a cross-media ontological framework, but this can also be done at the level of merging information at the feature levels of the distinct media involved. The best strategy still has to be discovered.

Some references:

[36] Download the presentation at: http://www.flarenet.eu/sites/default/files/Declerck_Presentation.pdf

Standards for Linguistic Annotation: ISO TC37/SC4: http://www.tc37sc4.org/ or the (past) Lirics project: http://lirics.loria.fr/

Standards for Multimedia Annotation: The MPEG-7 framework (an excellent overview is given in: http://gps-tsc.upc.es/imatge/_Philippe/demo/MPEG-7_Overview.pdf

Semantic Web: Standards for the Semantic Web are listed at W3C: http://www.w3.org. On the relation between Language Technology and Semantic, see also: http://semanticweb.dfki.de/Wiki.jsp?page=Main and http://ontoweb-lt.dfki.de/ (this one a bit outdated, but still valid from the content point of view).

S5 – Translation, Localisation, Multilingualism

Chair: Gerhard Budin - Rapporteur: Stelios Piperidis

Introduction

The cultural and linguistic diversity is one of the most important social assets of Europe. In addition to the 23 official languages of the European Union, there are dozens of minority languages spoken on the territories of EU member states plus hundreds of immigrant languages spoken in major European cities and regions. Local and national governments as well as European institutions are allocating significant financial resources to providing translation and interpreting services, multilingual information access and multi-media forms of cross-lingual information transfer. Localisation services have become a major industry branch and a professional service sector with growing market shares. Cross-cultural communication in most domains of science and technology, economy, art and culture, social affairs and other spheres of society requires the use of domain-specific terminologies in all languages concerned.

Language resources and language technologies (LRT) have become indispensable tools in order to enable translators, interpreters, technical writers, localisers, and other language professionals to provide high-quality services. Machine translation is increasingly used in industry, public administration, and other areas where millions of pages have to be translated every year. It is now also getting widely used by the grand public, thanks to the tools offered freely over the internet for document or message translation. And it starts being extended to the spoken language, with the perspectives of automatic interpretation of talks, courses and meetings, and the need to understand the huge amount of video now available on the Web. Computer-assisted translation methods such as translation memory systems, terminology management systems, localisation tools, etc. are widely used in SMEs and public services. Multilingual text corpora (aligned corpora, parallel texts, comparable corpora etc.) as well as multilingual lexical corpora, lexicons, term bases, etc. are being prepared and used for diverse application scenarios. Quality management tools such as translation metrics, standards for translation service providers, semi-automated workflow and project management systems are language technology applications that are increasingly used by language professionals.

Discussion, Objectives, FLaReNet Claims

The FLaReNet project aims at analysing the situation and the processes described above and at deriving a roadmap for initiating new research and development initiatives, coordination efforts needed to integrate existing LRT, and industrial innovation processes in order to enhance the competitive strength of European SMEs in translation and localisation industry, to support public services in providing high quality multilingual information and the education sector in its efforts to provide multilingual education and to enhance foreign language learning in the context of e-learning initiatives.

Enhancing multilingualism in Europe requires a concerted effort of all sectors and communities concerned: language technology providers, language resource producers, the users of LRT in their diverse application contexts, policy makers and other decision makers in public administration and other groups concerned should work together in order to better understand the changing and increasing needs of LRT users which is a pre-requisite to produce tailor-made LRT and technology-based language services that are based on realistic and sustainable business models.

The goal of this session is to identify urgent needs, assess current trends, and formulate concrete recommendations for further action in this strategic field.

Questions

5.1. What are the most important problems in the field of translation and localisation services and how can LRT help solve these problems?

5.2. What are the practical requirements that have to be fulfilled by LRT in order to be able to be relevant and helpful to translation and localisation services?

5.3. What are current users' needs when using LRT for translation and localisation services and are they met by available LRTs?

5.4. What are current trends in language technology developments for machine translation, computer-assisted translation, terminology management, localisation engineering, automatic interpretation and other multilingual technologies? Are there emerging paradigm shifts in these technologies?

5.5. What is the desired degree of automation? How much interaction with translation systems can users afford, if any?

5.6. How effective has the role of user feeback to automatic translation services offered by a number of search engines and systems been so far? Is this one of the possible viable solutions for systems improvement?

5.7. What is/can be the role of semantic web technologies in translation and localisation technologies?

5.8. How do we assess the usefulness, quality and relevance of available language resources that are used in translation and localisation technology development? Do we need new types of LRs?

5.9. How do we assess available multilingual information strategies, web portals, web services, cross-lingual search engines, automatic interpretation and other tools and resources enhancing real-life multilingualism?

5.10. How can LRTs be better used in the education sector in e-learning contexts, for language learning, MCLIL approaches (multilingual content and language integrated learning) and other challenges?

Language Resources and Tools for Machine Translation: Trends, Demands, Predictions[37]

Hans Uszkoreit, DFKI - Germany

[37] Download the presentation at: http://www.flarenet.eu/sites/default/files/Uszkoreit_Presentation.pdf

Outlook for Spoken Language Translation[38]

Marcello Federico - Fondazione Bruno Kessler
38100 Povo (Trento) – Italy
<surname>@fbk.eu

Introduction

For the European Union, with its 23 official languages and many more spoken languages, the availability of fast, reliable, and cheap translation and localization technologies is widely perceived as a strategic goal for thee neat future. This position paper provides a personal outlook for language resources and language technologies (LRT) related to the problem of spoken language translation (SLT), which combines the problems of automatic speech recognition and machine translation.

Tasks

While, e.g., broadcast news translation can be treated similarly to written text translation, different ideas of translation could be considered for conversational speech. For this task, humans professional translators typically refer to three ”interpreting modalities”: simultaneous, consecutive and *liason*. Simply speaking, all modalities require the human interpreter to listen to a given amount of speech, to recount what has been said, to listen again, and so on. Probably, the less ambitious scenario for SLT might be the one of **simultaneous interpreting**, which typically requires the human to translate at very short intervals, e.g. few seconds, or even in real-time. Besides being physically very demanding, simultaneous interpreters, due to the strict time constraints, are less able to exploit their linguistic and domain knowledge. Both reasons make users accept less fluent and almost close to literal translations.

Evaluation

Human and automatic evaluation of SLT should take into account important differences between written and spoken language. Practically, how should input sentences containing disfluencies and syntactic errors be treated? what kind of human translations should be taken as target references? The simultaneous interpreting scenario would suggest to put **more emphasis on adequacy** rather than fluency. Moreover, appropriate reference translations could be obtained by transcribing human interpreters working in realistic conditions.

Language Resources

The availability of language resources for MT has dramatically increased over the last decade, at least for a subset of relevant language. Unfortunately, the increase in quantity has not gone in parallel with an **increase in assortment**. Large parallel data are for the moment limited to texts produced by international organizations (European Parliament, United Nations, Canadian Hansard), press agencies, and technical manuals. Research on SLT needs bilingual and monolingual LRs from a wide spectrum of styles, genres, domains, topics, and situations, e.g.: conversations, lectures, speeches, documentaries,

[38] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Federico_Presentation.pdf

etc. Current performance of ASR suggest that a large amount of multilingual **LRs could be collected automatically**, by recording e.g. multiple channels from multilingual broadcasts or speeches provided with simultaneous interpretation (e.g. forums and conferences).

Infrastructure

The collection of large LRs should be carried out by non-profit entities with 100% funding and be recognized as a fundamental service to the HLT research community. Data collection is indeed the backbone of current HLT research! Quality and price of European LRs should be comparable to those distributed by the Linguistic Data Consortium.

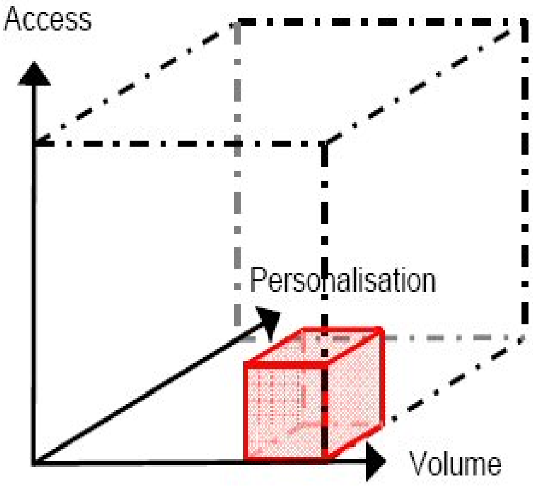
3 Challenges for Localisation[39]

Josef van Genabith - Centre for Next Generation Localisation (CNGL), Dublin City University, Ireland

Localisation is the industrial process of adapting digital content to culture, locale and linguistic environment, ideally at high quality, speed, volume and low cost. Localisation is a key enabling, value adding, multiplier component of the global manufacturing, services, software and content distribution industry. Currently, Localisation is facing three massive challenges:

> • **Volume:** the amount of content to be localised into ever more languages is growing rapidly and massively outstrips the supply of human translators.
> • **Access:** digital content delivery and interface devices are changing massively to enable pervasive, on
> the move and instant access to digital content.
> • **Personalisation:** the "raw material" for localisation is information. Information is most valuable when personalized to individual user requirements. Currently, however, Localisation operates according to a coarse-grained notion of linguistic, cultural and geographic locales.

Current state-of-the-art localisation technology focuses on basic off-line bulk corporate content localisation with low levels of personalization and traditional print or large screen/keyboard and desktop-based access modalities, to the exclusion of other localisation scenarios, and is unable to cope with the volume of (even "just") corporate content that needs to be localized.



The "Localisation Cube" and Current Localisation Technology

[39] Download the presentation at: http://www.flarenet.eu/sites/default/files/van_Genabith_Presentation.pdf

This and the combined effects of the three challenges result in massive opportunities missed: only a fraction of (corporate, institutional, Web) content is available across languages, there is a lack of support for instant access, new devices and interface modalities, and information composition and delivery is largely oblivious to user profiles, including important social and personal identities cutting across traditional linguistic and geographical boundaries.

- **Vision:** to enable people to interact with content, products and services in their own language, according to their own culture, and according to their own personal needs.

In order to achieve this vision we require technologies that address the triple challenges of Volume, Access and Personalisation: Language Technologies to address Volume and Access, and a novel combination of Adaptive Hypermedia and IR/IE technologies to address Personalisation. These technologies need to be integrated into the workflows of the Next Generation Localisation Factory (including Crowd Sourcing, pre- and post-editing, TMs, Terminology Management, Quality Control etc.). The factory will be virtual, self- configuring and, depending on a localisation request, optimally address each point in the space defined by the Localisation Cube (defined by the axes of Volume, Access and Personalisation), with configurable quality and speed.

Assessing User Satisfaction with Embedded MT[40]

Tony Hartley - Centre for Translation Studies – Leeds, UK

Evaluation should not only tell us *that* an MT system is getting better (or worse) but also inform us *how* – qualitative feedback is the motor for improvement. Sadly, progress this century has been impeded by an obsession with capturing quality in a 'magic number' and an ostrich-like attitude to the true characteristics of MT output. To reprise Church and Hovy (1993), we know we can live with some 'crumminess', but we still can't spell out the limits of this tolerable imperfection. We can rank systems against an arbitrary human gold standard, but we can't reliably predict from the linguistic make-up of their output how fit it will be for a given purpose.

In Call 4 the Commission has acknowledged that there are no absolute standards of translation quality in emphasising the need for quality indicators that measure fitness as 'communicative success in the particular use situation(s) being addressed'.

Defining quality relative to task performance and embedding MT into a flow of information has the advantage of clarifying the measure of success or failure but challenges us to adopt more diverse and finer-grained metrics than a text-based metric which simply assigns a number to the similarity between one 6-7k-word corpus and another.

Performance-based evaluation of MT – assessing the performance of humans (and other systems) using MT output to accomplish a specific task – precedes ALPAC. Among the more linguistically-oriented tasks that researchers have studied are: named entity recognition, co-reference chaining, information extraction, information retrieval, document triage and post-editing. Generally, these experiments have been used to arrive at a summative assessment of the capability of participating systems. But they could also be used formatively, if we could characterise just what degradation of linguistic context causes failure in identifying an NE or filling an IE template, for example. We need to revive task-oriented error analysis; it can benefit not only RBMT but also the growing paradigm of linguistically informed SMT.

To this end, we should see if it's possible to exploit the corpora and human annotations generated within such performance-based evaluation projects and also within corpus-based evaluation campaigns, such as CESTA, DARPA and NIST.

We should be aiming to do as revisers of human translation do – separate the review of TT well-formedness and fluency from the review of its semantic correspondence with the ST. If we can identify the structural properties of texts that promote or inhibit different defined end uses, then we can subsequently identify their presence or absence in a target-language text monolingually. There have already been reasonably encouraging attempts at ranking MT systems in this way, without reference to the ST. Factual equivalence can be assessed as a textual entailment task, which can 'deverbalize' both ST and TT. We can then experiment with different weightings of the results of the two exercises, as we do already with precision and recall.

[40] Download the presentation at:
http://www.flarenet.eu/sites/default/files/Hartley_Presentation.pdf

New tasks for MT, such as enabling social networking across language barriers, will drive the development of metrics for assessing the preservation of affect and attitude (and more) as users come to expect to use MT not only to convey facts but also to project social identity.

Institutional Translators and LRT[41]

Josep Bonet – European Commission, DG Translation, Luxembourg

LRT specialists and translators can be antagonists with different interests and needs. While the ultimate goal tends to be the same, the means to achieving it are not the same. The goal is to eventually produce better quality translations (within the limits of fitness-for-purpose) with the minimum effort possible; *i.e.*, to enhance productivity while maximising the quality inside the boundaries set by the needs expressed by the requester of translations. (This can be called ROI or work in a context of budgetary restrictions, depending on whether you are in the private of public sector). But while specialists try to use as many tools and resources as possible and want that resources are of the highest quality possible, translators tend to use as few tools as possible integrating many resources and prefer wider subject coverage rather than deeper quality levels.

Quality vs. Quantity: the answer is automation. In the 90s we had two lines of thought: those who preferred to only store confirmed translations which had received all kinds of blessings and those who archived all translations produced. The latter won the battle by getting critical mass way earlier. Cleaner resources, more structured information probably give better results, if you can pay for it. If the way to it is heavy use of human resources, it is better to abandon the idea. Do it with technology or rather forget. Just as Google freed us from the need to classify information, just not to be able to find it again, these things must be done by machines. Indeed, nothing compares to fully automating processes. When translation memory creation or TM clean-up was reduced to pressing a button, DGT entered the TM era.

Linked to this is the need to concentrate on content and forget about the rest. IT-aware translators do all kind of tasks which should be done by others. Why not the machine? On the other hand, contradictory as it may seem, translators tend to need to see the final product in context. Let's strip documents of all formatting while offering rendering capabilities to the user.

Interoperability and modularity. Who needs full-fledged, all-inclusive solutions? One-size-fits-all means no-one is well served. Modular approaches are better, allowing users to choose the modules needed from the best vendor. This means that they must be interoperable through open standards: LRs become reusable with any product available.

The interest of MT. In institutional translation, MT helps to control demand, diverting the part of it where less quality is needed or where the only need is to get the gist. For actual high-quality institutional translation, sub-segment phrase reuse looks more promising. It will be necessary to analyse translators' reaction to the texts proposed by the machine. Another issue is that institutional translation requires high internal consistency, which can only be guaranteed with TM and partly with RBMT, but hardly with SMT.

[41] Download the presentation at: http://www.flarenet.eu/sites/default/files/Bonet-Heras_Presentation.pdf

DGT has many resources available which are considered very useful, although it could be improved via automated processes, like language identification to correct wrong language tagging. Resources should be pooled in international repositories (like the TDA project). However, many organisations still have legal issues pending.

LRs available in translation organisations contain mainly translated texts, their quality being inferior to texts drafted originally in the target language. Better methods allowing to use monolingual original sources would be much welcomed.

Language Technology in the European Parliament's Directorate General for Translation. Facts, Problems and Visions[42]

Alexandros Poulis - ITS-DGTRAD, Luxembourg

Multilingualism is a core and indispensable aspect of the European Parliament (EP) and **translation** is the process that provides all Members of Parliament with equal access to the legislative procedure as far as their native language is concerned. At the same time, it is our duty to increase **transparency by** removing any language barriers that might discourage EU citizens from exercising their fundamental right of access to public information concerning the EPs legislative or other activities. Although the complexity and the workload increase with every new Member State, we have been able to meet most of these demands, but **at what cost**? What can Language Technology (LT) do to help us reduce that cost? What **problems** can be realistically addressed and what **visions** could lead us to the efficient implementation of LT in the **e-parliament**[43] era?

We will try to divide the problems or disadvantages directly or indirectly related to the translation procedure into three main categories: Workflow-related problems at an institutional level (among the various services of the EP), Workflow management harmonisation within the Directorate-General for Translation (DGTRAD) and problems related to the translation procedure per se.

If we were to look at the document life-cycle in the European Parliament we would realise that there are often technical gaps between the various services. There are cases where the authoring services do not take into account the abilities of our Computer Aided Translation (CAT) tools and produce documents with content that is either untranslatable (e.g. images) or cannot be processed by common CAT tools (e.g. embedded objects in word documents). Apart from the technical aspects there are also some problems of a linguistic nature in the same category, for which translation-oriented authoring might be a solution. For an organisation the size of the European Parliament, it is a major challenge to implement *LT solutions that bring major localisation processes such as authoring, translation and distribution under a common platform, monitoring each step of a project's life-cycle in an integrated workflow management environment*. The implementation of **standards** such as XML and **XLIFF** would be a big step towards a solution of the format incompatibility problems.

But how do workflow management and harmonisation relate to LT solutions? The fact is that the DGTRAD employs almost 1400 translators and assistants with various degrees of IT competence and representing different generations - since some of them have been working for the EP for more than 20 years. It is true that we cannot implement a common workflow for all language units unless the majority of their users are able to use the tools on offer with a cost-effective level of competence. The situation becomes even more complicated if we take into consideration the number of applications used by each translator, including dictionaries, terminology databases,

[42] Download the presentation at: http://www.flarenet.eu/sites/default/files/Poulis_Presentation.pdf

[43] **E-Parliament** is a legislature that is empowered to be more transparent, accessible and accountable through ICT. It empowers people, in all their diversity, to be more engaged in public life by providing higher-quality information and greater access to its parliamentary documents and activities. Furthermore, it is an organisation where connected stakeholders use information and communication technologies to support its primary functions of representation, law-making and scrutiny more effectively.
Through the application of modern technology and standards and the adoption of supportive policies, it fosters the development of an equitable and inclusive information society.

online resources etc. Ideally, all these applications should be available in one common, ergonomic user-interface.

User-friendliness affects the translation process as well. Most CAT tools are quite complicated from a translator's point of view, especially when it comes to procedures which are irrelevant to the linguistic task of translating a document such as repairing damaged tags, clean-up of application codes in the document, maintenance of translation memories and other application-specific aspects.

Even if a promising application appears on the market, it is extremely difficult to **evaluate** and offer it to the users since large-scale evaluation and subsequent deployment entail a high cost in terms of human resources. *It would therefore be very useful to have well-established standards for the comparative evaluation of LT solutions and more precisely CAT tools. Outsourcing the evaluation of Language Technology solutions is another option but we still have to overcome some barriers concerning the distribution of copyrighted materials such as our translation memory data.*

The same goes for other types of LT technologies such as statistical machine translation. The problem with SMT is two-fold. One possibility would be to evaluate SMT on the basis of open source and free software such as the baseline system offered by the ACL workshop on the SMT web-site - which presupposes additional, highly specialised human resources. Another possibility would be to allocate this task to a specialised SMT service provider but then we would have to cope with various legal matters related to the possibly classified nature of some parts of the corpus. *Concerning the first case, we would like to have user-friendly, platform-independent open source SMT software which would be easy to install, set-up and evaluate. On the other hand, a clear legal framework on the sharing of linguistic resources should be established so that institutions and companies are more willing to share their linguistic resources.*

The European Parliament is one of the few organisations required to translate multilingual documents. Each legislative (draft) proposal from the Commission is usually amended by more than one MEP. As a result, the amended document contains paragraphs or sometimes smaller chunks in various EU languages. These 'panaché' documents must then be translated into all official languages and today no CAT-Tool provides translation memories with multilingual source segments or the possibility of changing language pairs at will.

Another practical requirement for LT tools would be their **availability in all EU languages**: this is something which can be achieved either through the implementation of language-independent techniques or through language-specific tools covering as many languages as possible. At the same time, we believe that there should be a **common online access point for language technologies**. The number of language technology players is rapidly increasing and soon it will not be possible to follow up all the developments, especially those coming from academia, which are documented on different web-sites often only in the local language. The LT-World portal is a very good initiative in this direction and we look forward to more offers possibly, providing rss feeds including references to LT sites, products, communities and blogs.

“Cloud sourcing” for the translation industry[44]

Andrew Joscelyne - TAUS

The TAUS Data Association (TDA) is interested in innovative technologies, mindsets, and practices that drive the **translation industry** forward. This may or may not have an impact on or coincide with the agenda of the **translation research community**:

We believe that:

1. Everything linguistic (text and/or speech file from a computer or mobile phone etc) is a potential **language resource (LR)**. But not every LR needs to be owned, categorised and managed in some top-down way.
2. **Automated translation** + post-editing will become the norm for large sets of multilingual publishing, requiring massive bitext management beyond the reach of individual organisations.
3. The most appropriate vehicle for delivering massive bitext will be intelligent “**cloud sourcing**” **–** i.e. ensuring *quality* and “*trustability”* via dedicated resource clouds.
4. **Infrastructure** (technology platforms) must be separated conceptually and physically from LR creators, owners and users in the industry. Cloud + value-added services is more relevant to the future than lots of privately owned repositories.
5. **Collaboration** will become a vital *modus operandi* for various types of work. But as they have short attention spans, **communities** will be more responsive to shifting “buzz” than sustained business needs. We therefore need to experiment how the “open source” approach (which has proved its worth in computer programming and engineering disciplines) can be transferred to translation. Will translation turn into very large volumes of very small translated chunks, assembled and quality checked by machines?
6. **Simple, bottom-up per domain** is the most realistic way to kick-start LR building for the age of “cloud sourcing”, rather than trying to build complex top-down taxonomies and hierarchies. Machine learning may be preferable to committee work.

My brief talk will show how TAUS (Translation Automation User Society) has set up an open-access, dedicated “cloud” (TAUS Data Association) with a range of players to act as an “operating system” for the translation/localisation industry. It provides an “industrial” **model** that may be applicable to other LR actions. It is predicated on specialisation of purpose and collaborative intelligence.

[44] Download the presentation at: http://www.flarenet.eu/sites/default/files/Joscelyne_Presentation.pdf

S6 – Enhancing Market/Models for LRs: New Challenges, New Services

Chair: Khalid Choukri - Rapporteur: Jan Odijk

Introduction

The setup of distribution agencies, such as LDC (1992) and ELRA (1995), has triggered the establishment of a market place of LRs for language technology (LT) players with middlemen playing broker roles. Before that time, bilateral agreements were negotiated between HLT players and data producers e.g. terminology centers, dictionary publishers, corpus producers (mostly newspapers and the like).

The new landscape of the e-business confronts the market with new opportunities and challenges, while increasing the potential number of players.

We distinguish the market of LRs along two perspectives: (1) the market of making LR accessible and (2) the production of a LR market.

Discussion, Objectives, FLaReNet Claims

Market of making LRs accessible

Many important issues are to be brought forward:

- Structural aspects of the market of access to LRs (small players, global market). Focus should be put on the pricing policies that are considered by different players (producers, brokers, distributors, users), and the corresponding business model: pay per copy versus freeware (payment through public funds, info-commercials) versus registration/membership fee;
- Factors that have impact on the completion of the transactions, such as quick availability of the data, high quality documentation, consistent description and metadata, flexible licensing, quality of the data (object of discussion in other sessions);
- Profiles and business relationships between supply chain partners (producers, providers, distributors, and users);
- New trends of e-marketplace (mostly broker-managed online market), together with electronic licensing (electronic signatures), on-line payment, on-line data transfer issues, etc.;
- Emergence of “local players” for a given language or a country (e.g. TST-Central in the Netherlands) and the need for a partnership when it comes to selling worldwide.

A key issue in this market place is that most of the potential providers (LRs producers essentially interested by their technology development) are not very much interested in supplying such resources to third parties that may be their competitors. The lack of incentive for such suppliers to join the marketplace should be discussed in particular on how to balance the interests and the goals of the provider and those of the users. We may want to focus on the benefits of joining this marketplace, but also in terms of the possible consequences of not joining. The consequences have to be drawn also with respect to the funding potentials.

LR production market

The production business model should also be discussed. So far we have experimented a few models: individual productions paid for by private customers, some publicly (and partially) funded resources, possibility of a pooling of resources to produce and then share a set of similar resources (SpeechDat family), etc. It is important to discuss this LR production "market" as well.

Questions

6.1. How to foster the emergence/growth of a business for companies (e.g. publishers) with long-standing traditional economic and business models in the sector of HLT?
6.2. How to promote the emergence/growth of new business players that focus on the market to access LRs with their own economic and business models?
6.3. How can motivational factors be induced by FLaReNet in order for players to join the LR market place?
6.4. How can producers be encouraged to share and distribute their resources to third parties? What types of producers are more likely to be motivated?
6.5. The production processes are not harmonized to build on an economy of scale. Can production costs be rendered more sustainable?
6.6. Why the percentage of traded LRs is so low? How can we raise that?
6.7. Which new business models for LRs? (e.g. in accordance with new BM as creative commons and open licences that have its revenues from services such as "maintenance", conversion to in-house formats, evaluation, collection of different sources, etc.).

No resources without applications[45]

Gregor Thurmair – Linguatec, Germany

Creation of resources is a labour-intensive task (search, cleanup etc.). This task needs to be funded. Funding, if not in an academic context, requires that potential industrial customers / users can use the produced resources in applications which allow them to make a business with them.

The main current business applications with business relevance are speech transcription, and dialogue systems on the speech side, and proofing tools, translation tools, and search tools on the textual side.

From a commercial point, people invest in the creation of resources in cases where the applications can be sold successfully; in such cases, additional services like customisation, adaptations to vertical merkets and specific domains, extension to new languages etc. are required, and can be offered as a service by Language Resources producers.

The different business applications have a different status wrt market success; the demand in LRs varies accordingly.

The production of Language Resources itself can be organised in different ways, from cooperative open-source networks to profit-oriented business.

[45] Download the presentation at: http://www.flarenet.eu/sites/default/files/Thurmair_Presentation.pdf

Buy a License or Pay for Service?[46]

Gianni Lazzari (PERVOICE S.p.A., IT)

Today users or system integrators buy licenses when they want to use language technologies applications.

Since a decade e-business is rapidly expanding and e-services like transcription, translation, indexing, cross-lingual information extraction are becoming very appealing considering:

> - the reasonable reliability of the information infrastructure available today and even more in the future;
> - the cost reduction for the user (pay per use);
> - the possibility to centralize the maintenance and evolution of the language technologies components of the e-service.

But... the value of the business chain is strongly pending towards the service provider… and the language components are becoming profitable only with a high volume business.

[46] Download the presentation at: http://www.flarenet.eu/sites/default/files/Lazzari_Presentation.pdf

Enhancing HLT Market with Cooperative Services[47]

Gábor Prószéky - MorphoLogic

The number of players on the HLT market is decreasing, namely developers in countries with less elaborated languages cannot compete with general solutions of big players which, in many cases, are not as good for the language in question as theirs. Machine translation technology is a good example for this: MT solutions divide languages of the world into four groups: (1) English, (2) languages that are economically important, (3) other languages with MT solutions, (4) languages without MT.

Web-based MT services can be found for the above (1)-(2) translations for all (2) languages, (2)-(2) translations for some (2) languages, and (1)-(3) translations for a few (3) languages. Users would need good MT solutions for (2)-(2), (2)-(3) and (3)-(3) language pairs, but it is impossible without cooperation of the present players on then market.

According to the user tests, rule-based MT (RBMT) systems show a bit better results than statistical MT (SMT), but RBMT is expensive (in terms of both time and money). On the other hand, SMT needs much-much more data for languages in the above 3rd and 4th category. In principle, SMT systems offer a perfect solution, but their quality depends on the quantity of elaborated data, and we'll never have the same amount of data for the language pairs English-Spanish and Bulgarian-Danish.

If the main MT developers would cooperate and create a common service covering all the important languages (approx. 50) with the best machine translation solutions, then they had the chance to efficiently fight against solutions offering the same languages but with worse quality.

[47] Download the presentation at: http://www.flarenet.eu/sites/default/files/Proszeky_Presentation.pdf

Cheap or Expensive - what works?[48]

Gudrun Magnusdottir – ESTeam, Sweden

The current market trend of Google free translations has partly undermined the market of MT. HLT is expensive to develop and thus needs to have a price tag so that companies stay afloat. Current free solutions seem not to have an effect on the market of high end software. However, it does create problems for companies that provide low cost products for general purpose. What is the best way forward?

[48] Download the presentation at: http://www.flarenet.eu/sites/default/files/Magnusdóttir_Presentation.pdf

Speech-to-Text Solutions for the European Market - a SME view to language scalability[49]

Siegfried Kunzmann - European Media Laboratory GmbH, Germany

A brief introduction of the EML will be followed by our work on speech-to-text solutions for call centre communication, speech analytics, and voice messaging & searching. To bring these technologies to the European market and to support lots of languages we are trying to create a European SME partner network. To make the network working effectively across countries requires a strong focus on local business (as these markets are still small but predicted to grow rapidly) and a strong interest to collaborate across the network. The entry barrier for Speech-to-Text solutions is high – especially for SMEs – as access to leading-edge, robust technology (feasible through business contracts) and lots of real conversational data (expensive to produce) are required. Making large amounts of “real” conversational data, as a start for the official European languages, available for research as well as SMEs at affordable cost will ultimately drive wide adaption of speech(-to-text) technologies by lots of companies across Europe.

[49] Download the presentation at: http://www.flarenet.eu/sites/default/files/Kunzmann_Presentation.pdf