Open Grey for Natural Language Processing: a ride on the network

Gabriella Pardelli, Sara Goggi, Manuela Sassi
Istituto di Linguistica Computazionale, "*Antonio Zampolli*"
Consiglio Nazionale delle Ricerche (*CNR*)
Via G. Moruzzi 1, 56124 Pisa, Italia
[gabriella.pardelli, sara.goggi, manuela.sassi]@ilc.cnr.it

Abstract
The aim of this paper is to introduce the Open Access movement for Natural Language Processing (NLP) by means of a wide range of open access Grey Literature documentation available on the web. In 2008 Robert Dale, in the last issue of volume 35 of *Computational Linguistics* said: "There are a number of definitions of the term 'open access' in circulation, but almost all share the key principle that scientific literature should be freely available for all to read, download, copy, distribute, and use (with appropriate attribution) without restriction". At first glance it might seem that the Open Access movement has gradually become more influential in the field of language technology by building repositories accessible through the network. Today's digital archives are niches of intellectual production spread by means of a wide range of documents (such as journal articles and proceedings) which, paradoxically, the search engines do not always reach. The use of inappropriate terms in the formulation of queries and the fragmentation of repositories in this area of investigation does not allow to retrieve information on a large scale.
The full paper, after a first introductory section, will be organized in two sections: 1) the first dedicated to the methodology for searching and tracing open access resources and to the criteria for analyzing and selecting the online documentation; 2) the second devoted to a description of the state-of-the-art of Open Access Grey Literature material in a statistical and thematic scenario.
As things stand, standardization of computational systems interconnected by links and tools of various nature allowing Internet users to easily retrieve the information that the web naturally makes available would then be essential.
Topics: Sustainability, Public Accessible Resources, Product and Service enhancements, Open Access, Curation and Preservation
Keywords: Open Access Movement, Natural Language Processing

1. Introduction

Open Access is the key in the development of Information Society (IS), a new method for sharing scientific resources which influences the dynamics of creation and dissemination of knowledge. In order to share and spread this knowledge ever more sophisticated digital devices are tuned up while scientific institutions and associations are lately committed to the creation of dedicated repositories with the intent of giving wide visibility to their resources.
There is more: sharing open access information does not only mean retrieving objects of digital nature from the origins but also digitally reproducing source material from the far and recent past.

The definition of Open Access is rather tricky, as Merkel-Sobotta says in 2005: " 'Open access' means many different things to many different people. To use an example from US politics: it is as difficult to be anti-choice as it is to be anti-life. In the flux of ideas generated by the new and rapidly developing phenomenon of web publishing, open access proponents were able to convince others that traditional publishers were "anti-open access" or even anti-access, period. Tested against the realities of e-publishing, this did not last very long".
The following are a few examples of significant Open Access repositories of our field:

i) the Association for Computational Linguistics (ACL) built a rich repository called *ACL Anthology*, a digital archive of research papers in Computational Linguistics. This archive traces down the history of Computational Linguistics from first research of the '60s by retrieving and putting on the web the articles published in the proceedings of the most important international conferences of the field (i.e. COLING series).
ii) since some time several conferences publish their contributions as open access documents: the Global Wordnet Conference and the Language Resources and Evaluation Conference (LREC), just to make a couple of examples.
iii) *Machine Translation Archive* is an electronic repository and bibliography of articles, books and papers on several topics in the field of machine translation, computer translation systems and computer-based translation tools. This archive contains knowledge: its documents, accessible by everyone, provide an historical overview of automatic translation which might turn out to be very useful both for experts and non-experts of the field.

Notwithstanding the fact that conferences and workshops scientific material is widely spread on the web, the available search engines – though very sophisticated – are not nowadays able to provide a comprehensive plan of open access resources for Language Technology.
As a matter of fact, the first decade of the new millennium has suddenly witnessed first the growth and then the rapid increase of the so-called "social networks" which totally transformed the way information is transmitted: nowadays the World Wide Web looks like an enormous collection of documents inter-connected and linked to the various search engines by sharing the same paradigm (Web 2.0).
The selection, conservation and storage of digital content apparently makes the users' fruition easier : but is this assumption really true?
To formulate appropriate and effective queries for a search is a difficult task for users and requires a careful terminological selection for obtaining the most from an Information Retrieval system: *Information Retrieval* is the academic discipline which studies the methodologies, tools, techniques and languages for searching and retrieving relevant data for an information need.

2. Web search technology

The term Information Retrieval was introduced by Calvin Mooers in 1951, who defined it in this way: "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information. It is another, more general, name for the production of a demand bibliography. Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation. Information retrieval is crucial to documentation and organization of knowledge". (Mooers, 1951, p. 25).
The continuous development of the Internet-related technology makes available to the user a huge quantity of information perpetually increasing: "The Web is immense, free, and available by mouse click", as Adam Kilgarriff and Gregory Grefenstette said in 2003 (Kilgarriff and Grefenstette, 2003, p. 333). But though Internet unsettled the traditional scientific communication channels because publications on the web do not have (and do not need) any preliminary filter and sharing documents on the net becomes knowledge open to everyone, it is sometimes difficult to assess the quality of a document as well as to retrieve its semantics if the address of the portal is unknown to the user. Users are therefore often in difficulty when they submit a query and the web answers with a wide range of documents, most of them lacking any identification feature (such as place and time of creation) apart from the topic.

Today scientific documentation produced by the academia is transmitted thanks to the respective institutional web sites by means of ever more sophisticated software platforms for managing documentation; but, on the other side, there is a lack for what concerns a pertinent information retrieval: the answers to users' queries are usually not thorough and precise enough. When a user enters a query into a search engine (typically by using keywords), the engine examines the pre-existing indexes and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text; afterwards, documents are ranked according to their relevance probability and shown to the user in such an order. Given the results, the user can decide to refine the query and the whole cycle starts once again.
In this scenario, filtering the results of a query for selecting those valuable from a qualitative point of view amidst the great amount of useless information is a task which requires experience and patience. The classification system of the Web makes available to users several sophisticated search algorithms which do not have, though, an immediate impact on the human cognitive process of identification of the most significant and useful links for a given query. Therefore, for getting to what is really needed users have to start from a generic resource and afterwards follow the links generated by the first resource: search engines sound the web for capturing new pages by means of the URLs and then index them.
"Information retrieval today operates in a number of different contexts: full-text, digital libraries, the Web, online catalogs, and networked applications. IR utilizes a number of conceptual models such as: algorithmic, probabilistic Boolean, semantic, fuzzy logic, and vector space. As a result, IR has supported the creation of a number of different applications among them: latent semantic analysis, vector space analysis, information filtering, data mining, automatic indexing and classification, along with a number of paradigms for query formulation and information visualization" (Baeza-Yates & Ribeiro-Neto, 1999).

Whilst nowadays it is taken for granted that science concerns everyone thanks to the global net and in particular to the open access archives, it is also true that three centuries and a half have passed since the establishment of the journal called *Philosophical Transactions of the Royal Society of London* (1665), founded for the dissemination of scientific contributions.

3. Where to search for Open Access Language Technology documents?

Open Access documentation in Natural Language Processing, Computational Linguistics and Human Language Technology domains is not always easily retrievable although a massive amount of precious information has been published on the web by academia, association and private companies since many years.
In the '60s, the research in the field of natural language processing consolidated and consequently new associations were born and international journals were founded: here below a mention of some associations and the respective scientific open access production available on their portals:

- ≈ In the past: in 1959 the *Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée* (ATALA) was born; [in 1965, ATALA becomes the *Association pour le Traitement Automatique des Langues*]; today: http://www.atala.org/
  Grey Literature for Language Technology: online proceedings of the TALN conference (*Traitement Automatique des Langues Naturelles*) to be found at http://www.atala.org/-Conference-TALN-RECITAL

- ≈ In the past: in 1962 the *Association for Machine Translation and Computational Linguistics (AMTCL)* was founded [in 1968 it becomes the *Association for Computational Linguistics (ACL)]*; today: http://www.aclweb.org/index.php?option=com_frontpage&Itemid=1
  Grey Literature for Language Technology: online proceedings of the ACL conference series can be found on the ACL Anthology ("A Digital Archive of Research Papers in Computational Linguistics") website at http://aclweb.org/anthology-new/

- ≈ In the past: in1965 the *Association Internationale de Linguistique Appliquée* or *International Association of Applied Linguistics (AILA)* was established; today: http://www.aila.info/
  Grey Literature for Language Technology: NO open access material.

- ≈ In the past: in 1973 the *Association for Literary and Linguistic Computing (ALLC)* was born; today: it has a new name: *The European Association for Digital Humanities* (http://www.allc.org/)
  Grey Literature for Language Technology: online proceedings of the ALLC conferences can be found at: http://www.allc.org/conferences

- ≈ In the past: in 1978 the *Association for Computers and the Humanities (ACH*) was founded*;* today: http://www.ach.org/
  Grey Literature for Language Technology: the collection includes abstracts from the "Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities" and covers the years 1996, 1997, and 2000-2003: http://67.207.129.15:8080/dh-abstracts/search

- ≈ In the past: in 1991 the *European Association for Machine Translation (EAMT*) was born; today: http://www.eamt.org
  Grey Literature for Language Technology: 1) "Archive de la Traduction Automatique - Index des organisations": http://www.mt-archive.info/foreign/organisations-french.htm; 2) electronic repository and bibliography of articles, books and papers on topics in machine translation, computer translation systems, and computer-based translation tools: http://mt-archive.info/

- ≈ In the past: in 1995 the *European Language Resources Association (ELRA)* was established; today: http://www.elra.info/
  Grey Literature for Language Technology: online proceedings of the LREC (*Language Resources and Evaluation Conference*) series are at: http://www.elra.info/LREC-Conference.html

- ≈ In the past: in 2000 *The Global WordNet Association (GWA*) was born: today: http://www.globalwordnet.org/
  Grey Literature for Language Technology: online proceedings of the GWA conferences can be found at: http://www.globalwordnet.org/gwa/gwa_conferences.html

At last, a mention to the Grey Literature production of our Institute of Computational Linguistics stored on the PUMA "PUblication MAnagement System" Repository: http://puma.isti.cnr.it

3.1 The blind search

When a user tries to retrieve information on a given topic from online repositories there are several possibilities to formulate a query; for instance, given the query "Language Technology", the web replies with about 878.000.000 results in 0,26 seconds (September 25, 2012 at 4.30 p.m.):
Academic articles for language technology:

- ❑ of the state of the art in human language technology - Cole – Cited by 577
- ❑ Stirring up trouble about language, technology and …-Postman–Cited by 158
- ❑ Information extraction as a core language technology - Wilks – Cited by 69

1. ***Language technology*** **- Wikipedia, the free encyclopedia** *en.wikipedia.org/wiki/**Language_technology*** - Traduci questa pagina *Language technology* is often called human *language technology* (HLT) or natural language processing (NLP) and consists of computational linguistics (or CL) **...**
2. **Welcome to *Language Technology* World — LT World** *www.lt-world.org/* - Traduci questa pagina 4 Feb 2012 – A portal on the range of *technologies* that deal with human *language*. News, conferences, projects, organisations, systems, and resources.
3. **CELI - *Language* and Information *Technology*** *www.celi.it/* **...** l'analisi del linguaggio è diventata un fattore di successo dei nostri Clienti. I *Language* Specialists madrelingua di CELI lavorano in più di 30 lingue diverse.
4. ***Language* Learning & *Technology* - Home** *llt.msu.edu/* - Traduci questa pagina Online journal devoted to *technology* and *language* education research for foreign and second *language* educators. Full text of articles available.
5. **UNITN | Human *Language Technology* and Interfaces** *www.unitn.it/ateneo/.../human-**language-technology**-and-interfaces* Le Tecnologie del Linguaggio (Human *Language Technologies*, HLT) ci permettono oggi di interagire a voce con vari servizi automatici, ad esempio per **...**
6. [PDF] ***Language Technology* A First Overview - DFKI** *www.dfki.de/~hansu/LT.pdf* - Traduci questa pagina Formato file: PDF/Adobe Acrobat - Visualizzazione rapida di H Uszkoreit - Citato da 9 - Articoli correlati *Language technologies* are information technologies that are specialized for dealing **...** are also often subsumed under the term Human *Language Technology*.
7. **DFKI *Language Technology* lab** *www.dfki.de/lt/* - Traduci questa pagina Die Deutsche Forschungszentrum für Künstliche Intelligenz GmbH mit Sitz in Kaiserslautern und Saarbrücken ist auf dem Gebiet innovativer **...**
8. **CMU - *Language Technologies* Institute** *www.lti.cs.cmu.edu/* - Traduci questa pagina CMU/LTI offers MS and PhD programs in *Language* and Information *Technologies*.
9. ***Language Technology*** *www.lang-**tech**.org/* - Traduci questa pagina LangTech is the european forum dedicated to communities and organisations involved in the development, deployment and exploitation of *Language* and **...**
10. **Immagini relative a *language technology*** - Segnala immagini non appropriate
11. **FBK | Human *Language Technology*** *hlt.fbk.eu/* - Traduci questa pagina FBK - Fondazione Bruno Kessler. Human *Language Technology*. FBK > IT. News. 10 Sep 2012. Demo Paper accepted at ISWC 2012. 03 Sep 2012 **...**

From this generic query it is possible to retrieve 9 portals, 2 open access publications (Cole and Wilks), 1 review (Portman). But only 2 documents from this set satisfy the user……

3.2 The conscious search

Let's try with queries formulated by an expert user:

1 Query: ACL Anthology

- ♦ Query: Language Technology

Google custom search

About 11700 results (0,66 seconds)
1) <bold>**Language**, **Technology**, and Society Richard Sproat</bold **...**
File format: PDF/Adobe Acrobat
Language, **Technology**, and Society. Richard Sproat. (Oregon Health & Science University). Oxford: Oxford University Press, 2010, xiii+286 pp; hardbound, **...**
www.aclweb.org/anthology-new/J/J11/J11-1011.pdf

2) Draft for a road map on human **language technology**
File format: PDF/Adobe Acrobat
motto “How will language and speech technology be used in the information **...** Advances in human **language technology** will offer nearly universal access to on- **...**
www.aclweb.org/anthology-new/W/W02/W02-1302.pdf

3) Spoken **Language Technology**: Where Do We Go From Here?
File format: PDF/Adobe Acrobat
Spoken **Language Technology**: Where Do We Go From Here? Roger K. Moore. 20 20 Speech Ltd. Malvern, UK. Recent years have seen dramatic **...**
www.aclweb.org/anthology/P00-1003.pdf

4) Australasian **Language Technology** Association (ALTA)
The Australasian **Language Technology** Association (ALTA) was founded at the 5th Australasian Natural Language Processing Workshop, in Canberra, **...**
aclweb.org/anthology-new/docs/alta.html

5) Evangelising **Language Technology**: A Practically-Focussed **...**
File format: PDF/Adobe Acrobat
Evangelising **Language Technology**: A Practically-Focussed Undergraduate Program. Robert Dale, Diego Mollá Aliod and Rolf Schwitter. Centre for Language **...**
www.aclweb.org/anthology-new/W/W02/W02-0104.pdf

6) Letter to the Editor: **Language Technology** for Beginners
File format: PDF/Adobe Acrobat
Language Technology for Beginners. Ronald A. Cole 1. (University of Colorado). I am writing in response to Varol Akman's review (Computational Linguistics, **...**
www.aclweb.org/anthology/J99-4012

7) Does **Language Technology** Offer Anything to Small Languages?
File format: PDF/Adobe Acrobat
Does **Language Technology** Offer Anything to Small. Languages? Nick Thieberger. PARADISEC, University of Melbourne/. University of Hawai'i at Manoa **...**
aclweb.org/anthology-new/U/U07/U07-1002.pdf

8) PROJECTED GOVERNMENT NEEDS IN HUMAN **LANGUAGE ...**
File format: PDF/Adobe Acrobat
for human **language technology**, this paper will discuss the uses which will probably **... language technology** a suitable solution to maximize the effectiveness in **...**
aclweb.org/anthology-new/H/H93/H93-1056.pdf

9) Proceedings of the Australasian **Language Technology** Summer **...**
File format: PDF/Adobe Acrobat
and Australasian **Language Technology** Workshop (ALTW). 2003 **...** The Australasian **Language Technology** Association is proud to present its inaugural **...**
aclweb.org/anthology-new/U/U03/U03-1000.pdf

10) Human **Language Technology** can modernize writing and grammar **...**
File format: PDF/Adobe Acrobat
Human **Language Technology** can modernize writing and grammar instruction. Gerard Kempen. University of Leiden. P.O. Box 9555, 2300 RB Leiden, The **...**
aclweb.org/anthology-new/C/C96/C96-2172.pdf

Results: 10 open access Grey Literature documents found!
Given the fact that the enormous amount of data available on the web is difficult to query from a semantic point of view, the human interpretation is always needed - - but which are the assumptions/conditions for making an effective query?

Nowadays knowledge extraction can be performed in a satisfactory way if: i) the know-how of the state-of-the-art is updated; ii) there is a good skillfulness in navigating on the web portals; iii) there is the ability to interpret the data.

4. Conclusions

The web should be considered both as a knowledge repository and a knowledge dispenser: from this perspective, there is the need to create innovative paradigms for information retrieval, to establish features for semantic search on web portals as well as to achieve a certain degree of precision and recall, which are the coefficients measuring the performance of an information retrieval system:

- ✓ *Precision*: proportion of relevant data retrieved from the total data retrieved
- ✓ *Recall*: extent of relevant data retrieved from the total data relevant in the database.

These coefficients measure two different factors:

- ❑ *Noise* = non-relevant data retrieved;
- ❑ *Silence* = relevant data that have not been retrieved from the data base.

Retrieval models compute the degree to which certain elements answer to a query: a good model should be able to maximize recall and precision and minimize, respectively, "silence" and "noise".

References

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern information retrieval*. Reading, MA: Addison-Wesley

Dale R., LastWords What's the Future for Computational Linguistics? *International Journal for Computational Linguistics*, Volume 34, Number 4, 2008.

Kilgarriff A., Grefenstette G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics,* Volume 29, 3, Association for Computational Linguistics, 333-347.

Merkel-Sobotta E., Elsevier and Open Access, *Neuroinformatics*, 2005, Volume 3, Pages 5-9.
http://www.springerlink.com/content/x66772m265840111/fulltext.pdf

Mooers, C. N. (1951). *Making information retrieval pay*. Boston, Zator Co.

Ranger Sara L., Grey Literature in Special Library: Access and the Use*, Publishing Research Quarterly,* 2005, Volume 21, Number 1, Pages 53-63.

http://www.aclweb.org/anthology/J/J08/J08-4008.pdf
http://www.ling.ohio-state.edu/acl08/cfp.html
http://www.mt-archive.info/
http://www.aclweb.org/anthology/
http://www.cfilt.iitb.ac.in/gwc2010/
http://www.lrec-conf.org/proceedings/lrec2010/index.html
http://puma.isti.cnr.it//index.php?langver=it
http://www.greynet.org/greytextarchive.html
http://www.regione.emilia-romagna.it/wcm/LineeGuida/sezioni/generali/motori/documento_motori_di_ricerca.pdf
http://www.iva.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/information_retrieval.htm