UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

Publisher
UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by Novática <http://www.ati.es/novatica/>, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by Novática

UPGRADE was created in October 2000 by CEPIS and was first published by Novática and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

UPGRADE is the anchor point for UPENET (UPGRADE European NETwork), the network of CEPIS member societies' publications, that currently includes the following ones:
- inforewiew, magazine from the Serbian CEPIS society JISA
- Informatica, journal from the Slovenian CEPIS society SDI
- Informatik-Spektrum, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- ITNOW, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- Mondo Digitale, digital journal from the Italian CEPIS society AICA
- Novática, journal from the Spanish CEPIS society ATI
- OCG Journal, journal from the Austrian CEPIS society OCG
- Pliroforiki, journal from the Cyprus CEPIS society CCS
- Tölvumál, journal from the Icelandic CEPIS society ISIP

Cover page designed by Concha Arias-Pérez
"DNA in love" / © CEPIS 2010
Layout Design: François Louis Nicolet
Composition: Jorge Llácer-Gil de Ramales

Editorial correspondence: Llorenç Pagés-Casas <pages@ati.es>
Advertising correspondence: <novatica@ati.es>

UPGRADE Newslist available at
<http://www.upgrade-cepis.org/pages/editinfo.html#newslist>

ISSN 1684-5285

Monograph of next issue (August 2010)

"2010: Emerging Information Technologies (II)"

(The full schedule of **UP**GRADE is available at our website)

CEPIS UPGRADE
The European Journal for the Informatics Professional
http://www.upgrade-cepis.org

Vol. XI, issue No. 3, June 2010

Monograph: 2010 - Emerging Information Technologies (I)

(published jointly with Novática*)

Guest Editors: *Alonso Álvarez-García, Heinz Brüggemann, Víctor-Amadeo Bañuls-Silvera, and Gregorio Martín-Quetglas*

UPENET (UPGRADE European NETwork)

CEPIS NEWS

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

Trends in Natural Language Processing and Text Mining

Javier Pueyo and José-Antonio Quiles-Follana

Communication, information, opinions, and even feelings are shared, stored and encoded by humans and by their institutions in natural language (as opposed to the artificial, programming or structured languages used by computers). As linguists have experienced for centuries, the analysis and decoding of human language is a complex process, due to its pervasive, low precision, contextual, and ambiguous nature. The generalized use of computers and global networks of communication have caused most of our natural language exchanges (email, IM, reports, documentation, and even personal ideas, hobbies or stories) to be encoded and stored in digital format, and shared through computer systems. Natural Language Processing (NLP) techniques, developed in the field of computational linguistics, are certainly taking advantage of this fact, and are already being widely used in areas such as text mining, information retrieval, document clustering, opinion mining or knowledge management. In this article we take a closer look at external knowledge resources that are just now starting to be exploited to enhance and enrich NLP processes. We also analyze emerging uses of NLP procedures that combine the internal knowledge extracted from documents with the external or background information available to us, through specialized and structured data banks or semantic and conceptual dictionaries.

Keywords: Automatic Classification, Computational Linguistic, Document Clustering, Knowledge Management, Machine Learning, Natural Language Processing, Opinion Mining, Sentimental Analysis, Text Mining.

1 Introduction

Natural -as opposed to artificial- languages provide humans and their institutions (law agencies, policy and financial organizations, businesses, universities, hospitals, public administrations, industries, or citizen communities) with a sophisticated system to share information, facts, opinions, thoughts, judgments, beliefs, or feelings. For thousands of years, speakers have used the grammars and lexicons of their languages to encode their intended messages. As effective as they are for human communication, natural languages are lexically [1] and structurally [2] ambiguous, verbally and socially contextual [3], and even metaphoric [4] in nature, so although comprehensive descriptions of their phonology, morphology, syntax, semantics, pragmatics or vocabularies exist for many of them, we are still far from developing Artificial Intelligence (AI) systems that will automatically understand and produce language the way we do (for an updated AI overview, including a chapter on NLP, see for example [5]).

The field of Natural Language Processing (NLP) offers, though, diverse techniques that have proved to be very useful to automatically analyze, extract and provide knowledge from the intrinsically unstructured sources of information encoded in human languages. During the last 20 years NLP techniques have evolved from mostly symbolic rule-based (and Prolog/Lisp-programmed) methods, based on logic and linguistic introspection, that is, our own internal linguistic knowledge, to a more data intensive, statistical and probabilistic orientation to language processing (and also more programming language independent) based on the actual production and massive storage and manipulation of language.

Authors

Javier Pueyo is a researcher at the Institute of Language and Cultures (CSIC, Spain's Higher Council for Scientific Research). He was awarded a BA degree in Hispanic Languages and Literatures from the *Universidad de Deusto*, Spain (1990), a MA in Hispanic Linguistics from the University of Southern California, USA, (1993) and a PhD in Spanish Language and Literature from *Universidad de Deusto* (1996). His areas of interest include Linguistics, Sephardic Language and Literature, and development and exploitation of textual corpora and NLP software for the diachronic analysis of languages. He has recently joined the GSyC/LibreSoft research group (*Universidad Rey Juan Carlos*, Madrid, Spain) where he develops and applies NLP techniques to the analysis of Free Software and Free Culture projects. <javier.pueyo@gmail.com>.

José-Antonio Quiles-Follana holds a Telecommunications Engineer degree from the *Universidad Politécnica de Madrid*, Spain, since 1990. He worked in the Industry National Institute (INI, Spain) in several electronic war projects for the Spanish Department of Defense. Then he worked in the Bioengineering and Telemedicine Group of the *Universidad Politécnica de Madrid*, managing and developing research projects in the areas of medical imaging and telemedicine. Now he is a technological specialist in Telefónica I+D (Spain) where he has been worked in frecuency planning in mobile telecommunications, trouble ticketing and workflow management, fraud management in prepaid mobile telephony, knowledge management systems, human resources evaluation, web services infrastructure, rich client technologies, service oriented architectures (SOA), semantic web, automatic free text classification and conceptual search. <quiles@tid.es>.

2 Corpus Linguistics

The arrival of the Corpus Linguistics discipline in the late 60's [6] and the availability of annotated and catego-

rized linguistic corpora (such as the early one million-words Brown Corpus [7], the more recent 100 million-words BNC [8], ANC [9] or CREA [10], the syntactically parsed Penn Treebank [11], or the categorized Reuters collections [12]) allowed researchers to shift focus from working within a more linguistic theoretical framework to using massive linguistic information stored on computers, both for training and testing and for deployment of their models. The process of creating the first annotated corpora was basically manual, a painful but necessary step to develop automatic annotation tools. The existence of those manually collected and annotated corpora allowed the creation of the tools available nowadays to create, annotate, and analyze new corpora or document collections in many languages.

Reference corpora, intended to describe a whole language, and the tools developed from this kind of data, are too general to be applied to some specific areas. However, they are the foundation that ultimately allow the compilation of highly specialized — and probably much smaller — corpora for use in the areas in which specific businesses or institutions operate. Biomedical [13] or legal [14] collections are only two of the many encouraging examples that can be mentioned.

3 NLP Methods and Software

Tokenization, sentence segmentation, morphological tagging (or Point of Sale (PoS) tagging), stop-word removing, stemming, full lemmatization, and sentence chunking are some of the basic NLP processes applied to free-running text to provide it with some initial structure. More advanced techniques, such as syntactical parsing, anaphora resolution, named entity recognition (NER) and other kinds of semantic annotations are further applied to make most of the linguistic information contained in free-running text easily retrievable, and to somehow capture its meaning. Large and detailed rule-based grammars ([15], [16] or [17]), machine-learning systems using different models (i.e. Decision Trees, Expectation-Maximization (EM), Hidden Markov Model (HMM), Naïve Bayes or Bayes Networks or, more recently, Support Vector Machine (SVM) applied for example to PoS or parsing), and valuable lexical resources (such as WordNet [18] or the multilingual EuroWordNet [19], for example) are being used to further develop these methods and also to improve the necessary disambiguation steps for each process.

The increasing use of statistical and probabilistic models requires the gathering of detailed frequency figures and the use of different relevance measuring scores. Information Gain (IG), Term Frequency–Inverse Document Frequency (TF-IDF), Mutual Information (MI), chi-square distribution, t-score and z-score are some of the most commonly used significance measures. N-grams (i.e. words or sequences of words), stems or lemmas, sentences, named entities, PoS, and further linguistic information within the data, as well as any measurable attributes of these components (i.e. word or sentence length) or their relations (i.e. collocations, colligations) are the features that machine-learning algorithms are fed in order to build efficient tokenizers, PoS taggers, parsers or topic classifiers (for a more detailed description of these and other NLP methods, see for example [20], [21] and [22]). All these techniques have already been developed and improved for many languages and, depending on the nature of the data to be analyzed, results are getting closer to 100% accuracy for some of the most common NLP procedures.

Although out-of-the-box commercial software is already available to do many of the tasks needed for NLP treatment of free-running text, and NLP is one of the most active areas at the research labs of the software industry [23][24][25][26], one of the most promising trends in the NLP implementation field, as in many other areas of computing, comes from the Free-Libre/OpenSource Software (FLOSS) communities. Even if the FLOSS programs, and the data distributed with them at the moment, might seem incomplete, and in some cases the implementations fall behind those developed by cutting-edge research groups or the software industry, the flexibility to integrate NLP tools and freely deploy and distribute them, and the availability of the source code and data will allow other IT groups, from outside the NLP world, to take instant advantage of most of the NLP methods and to integrate them in very promising ways within their products, solutions, or research. Some FLOSS implementations of NLP techniques include: FreeLing, a suite consisting of an executable program and development libraries for language analysis services [27], Weka, a collection of machine learning algorithms for text mining [28], GATE, a suite for annotation and many other language processing tasks [29], or NLTK, a collection of python modules and linguistic data for development in NLP [30].

4 Areas of Increasing NLP Usage

As mentioned above, language is all around us, and every aspect of human interaction is surrounded by unstructured verbal expressions, so there are no limits to the potential areas of use of NLP techniques to automatically process information. Any application that deals in some way with free-running text can take advantage of them. Some of the areas and applications in which NLP usage is fully established or emerging are described in the following paragraphs [31].

Automatic document spelling is an area that is beginning to emerge in areas of business such as publishing and corporate legal departments. The currently existing manual spelling procedures would step automatic correction without any user intervention. To achieve this goal it is necessary for the system to understandthe semantics and concepts inside the text in order to make spelling decisions. In addition to spelling other automatic NLP corrections which are now possible include hyphenation, grammar and style checking.

There are other typical tasks in processing collections of documents such as automatic classification. Here we have a set of predefined categories or topics, and every document is assigned to one (or more) categories based solely

on its contents. Another typical application when processing large amounts of text documents is clustering. In this case, the system has not a set of categories defined a priori, but it attempts to discover these categories seeking similar documents and grouping them into groups of documents that share common characteristics (always based only on textual content). The extraction of summaries is another useful task especially in today's society where we have access to vast amounts of text and do not have the necessary time to process it. Automatic summarization could allow almost immediate selection or discarding of large amounts of information.

NLP techniques have helped in a remarkable way in tasks such as identifying the language of the texts, searching and retrieval of information, allowing much more precise searches and applying contextual linguistic processing techniques available today.

In today's society, where we can have access to so much knowledge without moving from our table, from a single computer with internet connection, there is demand for automatic translation of texts in different languages. In its most basic level, machine translation replaces the words of a source language with words of the target language. With the use of linguistic corpus, using techniques such as syntactical labeling, entities and concepts identification, more complex translations may be attempted. Those techniques allow a much smarter translation than the mere substitution of words.

Another application, that is beginning to be relevant, is the automatic creation of free text. Natural language generation is the process of building a natural language text to communicate a specific goal. To generate natural language text, we start from knowledge to be transmitted, then we must decide how to organize that information, and finally the question arises how to produce text, including the lexical entries and syntactic structures. Free text generation is beginning to be deployed in question/answer systems.

Throughout history, humans have used language not only to transmit knowledge, but also feelings and emotions. Automatic detection of feelings, within the free text written in natural language, requires semantic analysis to allow automated understanding of content, analysis and use in the form of new knowledge or as an aid to decision making processes. Within semantic analysis many problems that are the target of many research efforts today arise: a) resolution of anaphors or pronouns; b) word sense disambiguation (polysemy) depending on the context in which they appear; c) semantic roles, as the meaning of a sentence is not based solely on the words it contains, but also in the order, grouping and relationships between them.

Related to the analysis of feelings is opinion mining. This application is devoted to determine the author's attitude about a topic. Attitude can be either the evaluation, affective, or emotional communication that is intended when writing a text. The growth and availability of resources and social web sites where opinions are expressed (blogs, forums, eCommerce, product catalogs, etc.), give rise to new opportunities and challenges to get this information statistically and then to apply it to decision making.

Word sense disambiguation is another problem of great interest today. It seeks to identify what possible sense or senses one word takes (of all potential senses, polysemy) in a given sentence. Research has advanced strongly on this issue in the last decade, using various techniques: a) methods based on dictionaries like WordNet, b) supervised machine learning methods, in which a classifier is trained for each word in a corpus of manually annotated samples with all senses of the word, c) unsupervised learning methods that look for clusters of senses, thus deducing the different meanings of words. Today, the methods that are doing best are supervised learning algorithms, which are getting an accuracy of 90% in the English language.

An immediate application of semantic disambiguation is conceptual search engines. One problem with current keyword search engines is that they do not identify the meaning of those keywords. For example, when you ask a traditional search engine to find documents containing the word "bank", it will not distinguish between documents talking about financial institutions and other documents talking about river banks.

Many companies and institutions in different areas of knowledge are investing research and development resources into applications involving NLP. One of the most active areas we can mention is the biomedical field. NLP applications are currently being applied to research in the biological and biomedical domains (for example by applying NER technologies to identify protein or gene names, and by using the document collection and interpretation techniques described above to cope with the overwhelming literature produced in the field). Increasing NLP development is also being carried out in the clinical domain, particularly important for patients, since diagnosis and quality of care and treatment heavily depend on the patient records described in unstructured, free-text clinical reports (for a recent paper describing examples and NLP procedures see [32]). On the other hand, health insurance companies (and other areas in which evaluation of potential customers is a key point) are also starting to consider NLP techniques (see for example [33]).

Software engineering projects and research, typically more concerned with artificial languages and quantitative analysis of programming code, could also benefit from the analysis of the knowledge accumulated in the form of linguistic information during the software development cycle: mailing lists, documentation repositories, source code comments, issue tracking system (BTS) databases, or version control system (SCM) logs, will eventually be analyzed and interconnected to enhance next generation forges and to improve the quality of the software development process.

More traditional applications in businesses dealing with customer satisfaction analysis will benefit from the advances on text classification and clustering methods, and will significantly cut expenses in manual evaluation of massive and continuous surveys, survey analysis and classification.

5 Enhancing NLP Data: Web-as-corpus and Other External Resources

As research in Computational Linguistics and related disciplines continues to steadily develop, enhance, and test algorithms and models for machine learning, feature selection and other areas of NLP, a good place to keep track of research advances is the open repository of papers of the Association for Computational Linguistics (ACL) journal [34] and also the proceedings of academic conferences in the field such as, for example [35], [36], or [37].

However, in the following sections we would like to focus on new trends in the use of the NLP machinery available today or in the near future. The generalized use of computers and global networks of communication have resulted in most of our natural language exchanges (email, IM, reports, documentation, and even personal ideas, hobbies or stories) being encoded and stored in digital format, and shared through computer systems. With the availability of this massive amount of linguistic data we might speculate whether the era of painfully compiled static corpora (transcription of oral language, acquisition of textual materials, followed by the required typing, or scanning/OCR, and correction of data) has come to an end.

In 2006, one year after the first WACWCL [38] organized by the Special Interest Group of the ACL [39], the Google Research Blog published the following announcement [40]:

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of one trillion words from public Web pages. [...] That's why we decided to share this enormous dataset with everyone.

That very same year, a 24 GB (1,024,908,267,229 tokens) English "Web 1T 5-gram" corpus started to be distributed [41]. In 2009 -one year after the 4th WACWCL [42] (significantly entitled Can we beat Google?)- the 27.9 GB "1Web 1T 5-gram, 10 European Languages" corpus [43] was made publicly available. The 2009 data complemented the 2006 English-only previous version with Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish n-grams. Finally, the web-as-a-(multilingual)-corpus was becoming a reality.

NLP tools and techniques will certainly take advantage of these and other gigantic corpora which are being developed. Although annotation of this data is always possible and desirable, a shift to a "more words and less linguistic annotation" approach is gaining momentum, both from the corpus linguistic theory field, it is still worth reading the early and "prophetic" words of Sinclair in 1992: "*As size of corpora moves into the hundreds of millions ... analysis should be done in real time ... hold the text in raw format and analyze it fresh each time analysis is required*" [44], and also the statement from the Google research team, the leading industry in massive data collection and management [45]:

So, follow the data. Choose a representation that can use unsupervised learning on unlabeled data, which is so much more plentiful than labeled data. Represent all the data with a nonparametric model rather than trying to summarize it with a parametric model, because with very large data sources, the data holds a lot of detail. For natural language applications, trust that human language has already evolved words for the important concepts. See how far you can go by tying together the words that are already there, rather than by inventing new concepts with clusters of words. Now go out and gather some data, and see what it can do. (p. 12)

It would seem that real-time processing, and massive raw data, are the trends being proposed for NLP. This is a defensible approach for some of the tasks dealing with language, but enhancing the data with further external resources on a real-time basis seems to be also a realistic path that, we believe, NLP methods will follow in the future. In the rest of this section we take a closer look at other external knowledge resources that are just now starting to be exploited in order to enhance and enrich many of the NLP processes.

The above mentioned WordNets, or dictionaries of concepts, make possible the enhancement of lexical terms with the identification of their possible senses and their conceptual relations. However, the valuable but extreme linguistic detail of WordNet's synsets (sometimes a simple and common term is assigned to 20 or even more senses) makes it difficult to fully rely on them for semantic annotation, word sense disambiguation, or in more practical tasks like enhanced searches (by using synonyms / antonyms / meronyms / holonym / hypernyms / hyponyms) or multilingual expansion of queries. The integration of further external resources, such as top-level ontologies like [46] and [47], has already turned out to be a reality with projects like MEANING [48], a next generation, comprehensive and multilingual lexical knowledge repository, which is already available for use and integration with NLP methods and applications [49].

Future uses for NLP procedures will combine the internal knowledge extracted from documents (defined in a broad sense) with the external, background information available to us, not only through specialized data banks or semantic and conceptual-multilingual dictionaries, as the ones we have just mentioned, but also, perhaps more importantly, through the APIs available to many other global and typically unstructured sources of knowledge. Some examples of sources which are good candidates for integration with NLP systems, and which are mainly comprised of contents provided by individuals in their own words and languages, are the following:

■ Personal or specialized blogs (grouped by categories in blog directory services) usually labelled with topics or mood states. As they become more accessible through powerful APIs, which have already been implemented by some blog providers such as Yahoo [50] or Google [51], and even by some blog directories [52], integration into the NLP ecosystem will be possible. Some interesting research using blogs for sentiment/mood analysis is already available [53][54].

■ On-line news providers are the professional counterpart of bloggers. As in the case of blog providers, we can envision the development of APIs to automatically integrate categorized news (by topic, country, or genre) to expand tasks such as text classification or clustering. Although access to massive and updated data from general on-line news providers might be used to enhance every aspect of the NLP tools, financial news analyzed in real-time by expert NLP systems could be a priority for professional operators and customers investing real money [55].

■ Wikipedia already offers open access to millions of encyclopedic articles in dozens of different languages. Articles are not only categorized, but it is also fairly easy to find the same entry translated or adapted to many of the other Wikipedia languages. Besides providing complete downloads of the entire encyclopedia, MediaWiki (the software behind the Wikipedia) offers APIs to interact in real-time with its contents. In the near future we expect Wikipedia and its users to embrace semantic annotation [56] which will make it easier to extract unambiguous named entities (such as places, persons or organizations) for NLP usage.

■ Massive linguistic information will also be available from open repositories of books in the public domain. As described in [57] "Google Book Search is an ambitious program to make all the world's books discoverable online" (p. 1). Besides Google Books, the impressive project Gutenberg [58] offers over 100,000 titles to be downloaded, not only for reading, but more importantly for redistribution and, hence, for use in NLP development.

■ We can postulate other interesting resources for the future: social network sites, such as Facebook, MySpace or Twitter, as their posts become publicly searchable; and, of course, the use of APIs to access electronic commerce sites, such as Amazon, for product description and, more interestingly for NLP, the access to open comments, reviews and ratings posted by customers.

6 The Future of NLP Applications

Besides the treatment of external data outlined in the previous section, there are other areas in which NLP will play a substantial role:

■ Semantic and conceptual search engines. Even though search engines have become extremely accurate using traditional string queries and ranking the matching results, NLP methods are required for search engines to go beyond the keyword matching paradigm and get to really understand what users mean by their queries, that is to say, to understand and expand the meaning and relations of the words in a search query. NLP techniques are also required to preprocess the data to be presented as a query result to make sure its contents satisfy what the user meant at their search. WordNet and ontology mapping [59], as well as the integration of the external knowledge resources presented above, combined with advances in automatic language understanding algorithms, will help to improve the semantic awareness of search engines, and will also complement ranking mechanisms of the results.

■ Besides taking search engines to a new conceptual or semantic level, NLP procedures will lead the new trends in search results visualization: automatic summarization, automatic classification and clustering of relevant documents by topics, or semantically-enhanced word clouds [60]. These are some of the components we will be seeing in the results pages of future search engines. Systems will post-process retrieved documents and, by using available NLP summarization and simplification methods, will be able to extract only the relevant parts to a query. For an example of research in this direction applied to blogs see [61].

■ An obvious field in which enhanced NLP techniques are to be applied in the future is immediate linguistic assistance development, both for business and for private environments. The overwhelming and increasing volume of email interchange requires the development of integrated intelligent email filtering and organization procedures, going beyond the common subject, sender, or date filters provided by current email applications. Understanding the contents of email messages and relating those contents to our previous messages and to our own previous filtering decisions requires NLP-aware engines. In the same fashion, assisted email answering, or reply prediction alerts [62], as well as immediate retrieval and visualization of personal or related information, are good candidates for NLP integration.

■ Chinese, Spanish, English, Arabic, German or French language users in the world add up to around 2,000 million speakers. 94% of the world's languages are spoken by only 6% of the world's population [63]. Minority and endangered language users could paradoxically benefit from a digital globalized world, and from applying and integrating NLP technologies already available to them. Linguistic data collection is a previous but in many cases impossible step (in human, research, and financial resources) for the processing of minority languages. However, machine-translation from and to minority languages will benefit from the increasing availability of data in digital format (Asturian, Aragonese, Basque, Catalan, and Galician Wikipedias [64] are just examples of user-contributed linguistic data in a mainly Spanish linguistic environment).

Many of these contents are translations of data originally produced in languages such as English or Spanish, making it easier to apply NLP learning techniques to develop machine-translation applications for these languages. In return, integration of minority languages within NLP applications will allow for a more comprehensive gathering of cross-linguistic information. For an example of possible

integration trends for minority languages, involving FLOSS software solutions and NLP technologies, take a look at Golfiño [65], the first Galician grammar checker for OpenOffice.org, which uses the above mentioned FreeLing software for the PoS tagging phase.

7 Conclusion: The Future is moving

Finally, the popularization of mobile devices and the need of both getting and providing immediate and relevant information, will require further integration of geospatial and global positioning systems, the external knowledge sources analyzed above, and also the non-linguistic multimedia content from blogs, web pages, or sites such as Wikipedia, Flickr, Picasaand YouTube. Multimedia content, being typically surrounded by free-running text [66], is easier to identify and analyze taking into account its linguistic context than relying only on audio/video/picture identification systems in order to incorporate them into applications providing enriched multimedia information to users. A good example of this kind of integrated systems can be found in the recent development of LibreGeoSocial [67], a FLOSS social network with a mobile Augmented Reality interface. Future integration of NLP techniques into this ecosystem is probably, in our opinion, the key to making the emerging augmented reality applications not only a "true" reality, but also a meaningful one.

References

[1] D. A. Cruse. Lexical Semantics. Cambridge [Cambridgeshire]/New York: Cambridge University Press, 1986.

[2] Kenneth Church, Ramesh Patil. "Coping with syntactic ambiguity or how to put the block in the box on the table". American Journal of Computational Linguistics, 8:139–149, 1982.

[3] Helen Leckie-Tarry, David Birch (ed.). Language and Context: A Functional Linguistic Theory of Register. London/New York: Pinter Publishers, 1995.

[4] George Lakoff, Mark Johnson. Metaphors We Live By. University of Chicago Press, 1980.

[5] Stuart Russell, Peter Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall, 2009.

[6] Henry Kucera, W. Nelson Francis. Computational Analysis of Present-Day American English. Providence, Rhode Island: Brown University Press, 1967.

[7] The Brown Corpus. <http://khnt.hit.uib.no/icame/manuals/brown/index.htm>.

[8] The British National Corpus (BNC). <http://sara.natcorp.ox.ac.uk/>.

[9] The American National Corpus (ANC). <http://american nationalcorpus.org/>.

[10] Corpus de Referencia del Español Actual (CREA). <http://corpus.rae.es/creanet.html>.

[11] The Penn Treebank Project. <http://www.cis.upenn.edu/~treebank/>.

[12] The Reuters Corpora. <http://trec.nist.gov/data/reuters/reuters.html>.

[13] Biomedical corpora. <http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml>.

[14] The Juris Corpus. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98T32>.

[15] LFG Pargram Project. <http://www2.parc.com/istl/groups/nltt/pargram/>.

[16] HPSG LinGO Matrix framework. <http://www.delph-in.net/matrix/>.

[17] XTAG Project. <http://www.cis.upenn.edu/~xtag/>.

[18] Christiane Fellbaum (ed). WordNet: An Electronic Lexical Database. MIT Press, 1988. <http://wordnet.princeton.edu/>.

[19] EuroWordNet. <http://www.illc.uva.nl/EuroWordNet/>.

[20] Christopher D. Manning, Hinrich Schuetze. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts. The MIT Press, 1999.

[21] Daniel Jurafsky, James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Second Edition). Prentice Hall, 2000.

[22] Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009.

[23] Google Research. <http://research.google.com/about.html>.

[24] Microsoft Research. <http://research.microsoft.com/en-us/groups/nlp/>.

[25] IBM Research. <http://domino.research.ibm.com/comm/research.nsf/pages/r.nlp.html>.

[26] Yahoo! Research. <http://research.yahoo.com/>.

[27] FreeLing. <http://www.lsi.upc.edu/~nlp/freeling/>.

[28] Weka. <http://www.cs.waikato.ac.nz/ml/weka/>.

[29] GATE. <http://gate.ac.uk/>.

[30] NLTK. <http://www.nltk.org/>.

[31] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, 2008.

[32] Danielle L. Mowery MS, Henk Harkema PhD, John N. Dowling MS MD, Jonathan L. Lustgarten PhD, Wendy W. Chapman PhD. "Distinguishing Historical from Current Problems in Clinical Reports—Which Textual Features Help?". Proceedings of the Workshop on BioNLP, 2009, pp. 10-18.

[33] Barry Glasgow, Alan Mandell, Dan Binney, Lila Ghemri, David Fisher. "MITA: An Information-Extraction Approach to the Analysis of Free-From Text in Life Insurance Applications", AI Magazine 19(1) (Spring 1998), pp. 59-72.

[34] Computational Linguistics. <http://www. mitpress journals.org/loi/coli>. Official Journal of the Association for Computational Linguistics <http://www.aclweb.org/>.

[35] Conference on Computational Linguistics (COLING). <http://www.coling-2010.org/>.

[36] Empirical Methods on Natural Language Processing (EMNLP). <http://conferences.inf.ed.ac.uk/emnlp09/>.

[37] International Conference on Machine Learning (ICML). <http://www.cs.mcgill.ca/~icml2009/>.
[38] Web as Corpus Workshop at Corpus Linguistics. <http://sslmit.unibo.it/~baroni/web_as_corpus_cl05.html>, 2005.
[39] SIGWAC. <http://www.sigwac.org.uk/>.
[40] Google Research Blog. <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>.
[41] Thorsten Brants, Alex Franz. Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia, 2006. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>.
[42] Stefan Evert, Adam Kilgarriff, Serge Sharoff. Proceedings of the
4th Web as CorpusWorkshop (WAC-4). <http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf>.
[43] Thorsten Brants, Alex Franz. Web 1T 5-gram, 10 European Languages. Version 1. Linguistic Data Consortium, Philadelphia, 2009: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>.
[44] J. Sinclair. "The automatic analysis of corpora". En J. Svartvik (ed.) Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82). Berlin: Mouton de Gruyter. (pp. 382-384), 1992.
[45] Alon Halevy, Peter Norvig, Fernando Pereira. "The Unreasonable Effectiveness of Data". IEEE Intelligent Systems, March/April 2009.
[46] A. Alonge, F. Bertagna, L. Bloksma, S. Climent, W. Peters, H. Rodríguez, A. Roventini, P. Vossen. Top Concept Ontology. "The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology". En Piek Vossen (ed.) Euro WordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht, 1998.
[47] Suggested Upper Merged Ontology (SUMO). <http://www.ontologyportal.org/>.
[48] MEANING. <http://www.lsi.upc.es/~nlp/meaning/>.
[49] Multilingual Central Repositor (MRC). <http://www.lsi.upc.es/~nlp/meaning/downloads.html>.
[50] Yahoo!'s Livejournal API. <http://developer. yahoo.com/>.
[51] Google's Blogger API. <http://code.google.com/intl/en/apis/blogger/>.
[52] BlogCatalog API. <http://www.programmableweb.com/api/blogcatalog>.
[53] Gilad Mishne. "Experiments with Mood Classification in Blog Posts". Stylistic Analysis of Text for Information Access, 2005.
[54] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs". International Conference on Weblogs and Social Media, ICWSM 2007 (Boulder, Colorado, USA).
[55] M. Costantino, R. G. Morgan, R. J. Collingham, R. Carigliano. "Natural language processing and information extraction: qualitative analysis of financial news articles". Computational Intelligence for Financial Engineering (CIFEr). Proceedings of the IEEE/IAFE 1997. pp. 116-122.
[56] Semantic MediaWiki. <http://semantic-mediawiki.org/wiki/Semantic_MediaWiki>.
[57] Luc Vincent. "Google Book Search: Document Understanding on a Massive Scale". Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, 2007. pp. 819-823.
[58] Project Gutenberg. <http://www.gutenberg.org/wiki/Main_Page>.
[59] Dario Bonino, Fulvio Corno, Laura Farinetti, Alessio Bosca. "Ontology Driven Semantic Search". WSEAS Transaction on Information Science and Application 1 (6) (2004) pp. 1597-1605.
[60] Byron Y-L. Kuo, Thomas Hentrich, Benjamin M. Good, Mark D. Wilkinson. "Tag Clouds for Summarizing Web Search Results". WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
[61] Michel Généreux. "Summarizing a Blog Search Engine Hits". Workshop on Web Search Result Summarization and Presentation, 2009, Madrid, Spain.
[62] Mark Dredze, Tova Brooks, Josh Carroll, Joshua Magarick, John Blitzer, Fernando Pereira. "Intelligent Email: Reply and Attachment Prediction". IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.
[63] M. Paul Lewis (ed). Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International, 2009. <http://www.ethnologue.com/>.
[64] Wikipedia. Asturian <http://ast.wikipedia.org/wiki/>, Aragonian <http://an.wikipedia.org/wiki/>, Basque <http://eu.wikipedia.org/wiki/>, Catalan <http://ca.wikipedia.org/wiki/>, Galician <http://gl.wikipedia.org/wiki/>.
[65] Golfiño. Galician spelling tool for OpenOffice.org. <http://www.imaxin.com/ficha.asp?IDproxecto=68>.
[66] Hrishikesh Aradhye, George Toderici, Jay Yagnik. "Video2Text: Learning to Annotate Video Content", ICDM Workshop on Internet Multimedia Mining 2009.
[67] LibreGeoSocial, GSyC/LibreSoft (URJC). <http://libregeosocial.morfeo-project.org/>.