COMPUTATIONAL LINGUISTICS

Nicoletta Calzolari
Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche (ILC-CNR), Pisa, Italy

Keywords: Computational linguistics, text and speech processing, language technology, language resources.

Contents

Summary

A brief overview of the field of Computational Linguistics is given. After a few sketches on the short history of the field, we focus on the area of Natural Language Processing providing an outline of some of the main components of computational linguistics systems. It is stressed the important role acquired by Language Resources, such as corpora and lexicons, in the last years. They are the prerequisite and the critical factor for the emergence and the consolidation of the *data-driven* and statistical approaches, which became undoubtedly predominant in the last decade, and are still the prevailing trend in human language technology. We point at the fact that linguistic technologies – for analysis, representation, access, acquisition, management of textual information – are more and more used for applications such as: (semi-)automatic translation, summarization, information retrieval and cross-lingual information retrieval, information extraction, question answering, classification of documents, search engines on the web, text/data mining, decision support systems, etc. The ever increasing necessity to access multiple types of information contained in natural language texts and documents stored in digital format (on the web or not) does give a strong stimulus to their development in today information and knowledge society. The language being

indeed the privileged means of interaction for social, economic and cultural activities, linguistic technologies are increasingly characterized as 'pervasive' or 'horizontal' technologies. They are in fact used in systems belonging to a wide range of applications and to various types of services on the web, often of multilingual nature: e-government, e-learning, e-commerce, e-business, e-culture, etc. Finally a few words on future perspectives of the field.

1. What is Computational Linguistics?

From a terminological point of view, various terms – and their acronyms – are frequently used for describing the field of Computational Linguistics (CL): besides the classical NLP (Natural Language Processing), over the years other terms have been employed, among which I mention at least LE (Language Engineering), LI (Language Industry) and more recently HLT (Human Language Technologies) or simply LT (Language Technologies) (*Or also TAL in French (Traitment Automatique de la Langue) and Italian (Trattamento Automatico della Lingua).*).

The term CL includes the disciplines dealing with models, methods, technologies, systems and applications concerning the automatic processing of a language, both spoken and written. CL therefore includes both Speech Processing (SP) or processing of the spoken word, and Natural Language Processing (NLP) or text processing. SP and NLP have closely linked objectives such as human-machine vocal interaction and human language understanding, to be used in many applications, such as machine translation, speech-to-speech translation, information retrieval, and so on.

Speech Processing – as the study of speech signals and the processing methods of these signals – deals with the human capacity of communicating orally and includes: encoding of the vocal signal; speech synthesis, that is creating a 'machine' able to generate speech and e.g. to read any text; speech recognition, that is a machine able to recognize spoken utterances; speaker recognition, to recognize the identity of the speaker.

Natural Language Processing deals with the human ability of understanding a language, and includes – as regards components and methods used: syntactic and semantic analyzers based on algorithms or statistics, models of knowledge representation based e.g. on dictionaries and encyclopedias, methodologies for automatic learning, and annotation and classification techniques as the starting point for information retrieval, information extraction, question answering, etc.; while – as regards applications – apart from machine translation (which has an essential role in the multilingual Europe), there are themes such as dialogue management, summary production, web search engines and knowledge management.Given that we cannot deal with the whole spectrum of areas covered by CL, we choose to focus this article on NLP and in particular on the fundamental role that in the last years the area of the so-called Language Resources have acquired in the field.

1.1. A few sketches in the history of Computational Linguistics

CL is a relatively young discipline, with a short history. The birth of CL coincides with

the invention and first usage of computers in the 40's and early 50's. This powerful tool aroused, during the Second World War, the interest of many scientists who foresaw the possibility of automatically translating Russian into English, just as Fortran could be translated into the machine language. This hope fostered the studies in the field of text processing, thus giving rise to a technology which has many applications in today Information Society. NLP actually developed from the pioneering Machine Translation (MT) – a very ambitious objective in particular for the times –, and started applying formal models – at the time worked out by generative-transformational linguistics schools – to the analysis, recognition and representation of the linguistic structures of texts or, vice versa, to the generation of texts starting right from the representation of these structures.

In the 50's also speech processing developed: e.g. a system for recognizing the numbers called out by a speaker was created in the Bell Laboratories.

CL technologies started being studied in various areas and by different communities: by computer scientists dealing with automatic translation, by telecommunication engineers interested in speech processing, by linguists studying the structure and the evolution of language, by cognitive psychologists concerned with the mechanisms of understanding. Since its beginning CL is thus established as a truly multidisciplinary area of research, with the peculiarity of being interdisciplinary between the usually far away areas of the Humanities on one side and hard sciences (information or computer science, mathematics, engineering, etc.) on the other.

As far as text processing is concerned, when at the end of the 40's electronic techniques started being applied to the processing of linguistic data, almost immediately two independent tendencies with very little in common developed: on the one hand machine translation and on the other hand lexical and textual analysis (creation of indexes, concordances, frequencies, etc.) mostly for literary computing.

As an example, in the 50's-60's in Italy the pioneering work of Father Roberto Busa, who succeeded in producing the automatic compilation of concordances in the complete works of Saint Thomas Aquinas or those attributed to him (up to a total of 10 million occurrences) at the Centre for Automatic Linguistic Analysis (CAAL) in Gallarate, should be mentioned.

In the same years, Antonio Zampolli, who moved his first steps in the field next to Father Busa, started broadening the horizons of text processing to wider aspects than the creation of indexes and concordances of texts, setting up a project (1969-70) – in collaboration with the Italian Chamber of Deputees – on *information retrieval* in the area of legislation, creating and using a large Machine Dictionary of Italian: a very innovative and visionary enterprise at the time.

As for Machine Translation, quoting John Hutchins: "In 1964 the government sponsors of MT in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to examine the prospects. In its famous 1966 report it concluded that MT was slower, less accurate and twice as expensive as human translation and that 'there is no immediate or predictable prospect of useful machine translation.' It saw no

need for further investment in MT research; instead it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support of basic research in computational linguistics. The ALPAC report was widely condemned as narrow, biased and shortsighted. It is true that it failed to recognize, for example, that revision of manually produced translations is essential for high quality, and it was unfair to criticize MT for needing to post-edit output. It may also have misjudged the economics of computer-based translation, but large-scale support of current approaches could not continue. The influence of the ALPAC report was profound. It brought a virtual end to MT research in the USA for over a decade and MT was for many years perceived as a complete failure."

MT research continued in other countries, and for example in Europe in the 70's the EUROTRA project for automatic translation was launched: it never reached its goal – the creation of a multilingual translation system – but it did contribute to the establishment of a broad network of European centers of Computational Linguistics and stimulated important novelties and experiences in the area of NLP. In many European countries the main centers of CL were born because of EUROTRA, which thus was instrumental to forming a quite new generation of computational linguists.

The growth of Computational Linguistics in the 70's was greatly due also to the interest in NLP shown by vast areas of the so-called Artificial Intelligence, a science aiming at developing methods and tools able to deeply "understand" human language but necessarily limiting itself to restricted linguistic fragments and to so-called "toy-systems", applying deep knowledge to very limited domains.

In the 80's one of the main issues was for sure the model called "Hidden Markov Model" (HMM), which is still currently applied to speech recognition and has been also widely used in NLP applications, such as morpho-syntactic analysis systems (taggers).
It was only in the 80's that the great importance of language resources (large lexicons and corpora) for the field to grow has been finally recognized: the first example in Europe of an international research project dealing with large-scale lexicons was the ACQUILEX project. ACQUILEX *(Funded within the European ESPRIT Basic Research Actions programme from 1988-94, as two subsequent projects.)* – where natural language definitions in printed dictionaries were the text to be analyzed for automatically acquiring syntactic and semantic information – constituted an essential step towards developing methodologies for automatic acquisition, but also design and representation of large computational lexicons. This process was referred to at the time as going from machine readable dictionaries (MRD) to lexical databases (LDB).

Language resources were not conceived as an end in themselves, but as an essential component to develop robust applications, and it was clear that they were the prerequisite and the critical factor for the emergence and the consolidation of the *data-driven* and statistical approaches, which became undoubtedly predominant in the last decade, and are still the prevailing trend in human language technology.

2. Automatic text processing

The automatic processing of written language can roughly be divided in two areas: the

generation of a text (synthesis) and its understanding (analysis).

By generation of a text is meant the creation of a text according to a set of concepts which have to be expressed following the rules of the language in which it is written. Examples of applications could be the generation of answers in a human-machine dialogue, the translation from another language or the creation of an article or book summary.

By understanding a text is meant the annotation or extraction of its conceptual content following phonetic, grammatical, syntactic, semantic and pragmatic or contextual rules, and/or using statistical processes. In this case applications are complementary to those already mentioned, that is the comprehension of sentences pronounced by the speaker in the human-machine interface and the comprehension of the text in the original language in order to produce the translation or generate a summary. Some level of linguistic analysis for "understanding" a text is used in information retrieval and information extraction systems as well as in proofreading tools (lexical, grammatical, syntactical and stylistic), and is practically necessary in any application having to deal with texts.

From the point of view of the research areas, the components and the infrastructures needed for building application systems, we should make reference to the systems of analysis related to the various linguistic levels (grammatical, syntactic, semantic), to computational lexicons (that is, vocabularies storing the knowledge about words necessary for analysis and generation), to ontologies, to statistical and machine-learning methodologies, to large annotated textual corpora (treebanks), etc.

From the point of view of applications, some amount of NLP is daily used in proofreading writing systems, in bilingual translations, in creating human-machine interfaces based on dialogue and in information retrieval, and so on.

-
-
-

Bibliography

Allen, J. F. (1995). *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings. [A comprehensive, in-depth description of the theories and techniques used in the field of natural language understanding.]

ALPAC (1966). *Languages and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C. 124 pp. [The famous report which came to the result that there was no need for further support of research and development in machine translation and that it would make much more sense to spend money for improving the quality of traditional translation by human translators.]

Bird, S. (ed). *ACL Anthology, A Digital Archive of Research Papers in Computational Linguistics*. http://www.aclweb.org/anthology/E91-1001. [A Digital Archive of Research Papers in Computational Linguistics from CL Journal and major Conferences such as ACL and COLING.]

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA, MIT Press. [The book, after a brief description of the classic artificial intelligence approach to NLP, presents a few definitions from probability theory and information theory, then introduces hidden Markov models and probabilistic context-free grammars, and discusses advanced topics in statistical language learning, such as grammar induction, syntactic disambiguation , word clustering, and word sense disambiguation.]

Hutchins, W.J. (1995). *Machine Translation: a brief history*. In E.F.K. Koerner, R.E. Asher. *Concise history of the language sciences: from the Sumerians to the cognitivists*. Oxford, Pergamon Press, pp. 431-445. [An historical perspective on MT, mentioning the major and most significant systems and projects.]

Hutchins, W.J., Somers. H. L. (1992). *An introduction to machine translation.* London, Academic Press. [A comprehensive introductory textbook for MT.]

Jurafsky, D., Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall. [A comprehensive introduction to natural language and speech processing.]

Maegaard B. et al. (2003). *Benchmarking HLT progress in Europe*. HOPE, Copenhagen. [The final report of the Euromap/Hope European project, outlining the situation of HLT in Europe.]

Manning, C., Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MA, MIT Press. [Thorough introduction to statistical approaches to natural language processing.]

Mitkov, R. (ed.). (2003). *The Oxford Handbook of Computational Linguistics*. Oxford Handbooks in Linguistics, Oxford University Press. [A reference book that provides a wealth of information on computational linguistics and natural language processing.]

Varile G.B., Zampolli A. (Managing Editors). (1997). *Survey of the State of the Art in Human Language Technology. Linguistica Computazionale*, XII-XIII. Giardini Editori e Stampatori, Pisa and Cambridge University Press. [The book, sponsored by the Directorate General XIII of the European Union and the Information Science and Engineering Directorate of the National Science Foundation, USA, offers the first comprehensive overview of the human language technology field.]

Walker, D., Zampolli, A., Calzolari, N. (eds.). (1995). *Automating the Lexicon: Research and Practice in a Multilingual Environment.* Clarendon Press, OUP, Oxford. [The book, based on the groundbreaking Grosseto Workshop, provides a baseline and a reference point on many issues associated with computational lexicons.]

Wilks, Y., Slator, B., Guthrie, L. (1996). *Electric Words: dictionaries, computers and meanings*. Cambridge, MA: MIT Press. [A general survey of and introduction to the entire range of work in lexical linguistics and corpora -- the study of such on-line resources as dictionaries and other texts -- in the broader fields of natural-language processing and artificial intelligence.]

Zampolli, A., Calzolari, N., Palmer, M. (eds.). (1994). *Current Issues in Computational Linguistics: in Honour of Don Walker*. Linguistica Computazionale, Vol. IX-X, Giardini Editori, Pisa and Kluwer Academic Publisher, Norwell (MA). [The widespread interests and interdisciplinary approach of Don Walker are mirrored in the rich and diverse collection of papers in this volume which portray many different approaches to Computational Linguistics.]

Biographical Sketch

Nicoletta Calzolari was born in Ferrara (Italy) in 1947 and graduated in Philosophy in Bologna University in 1969.

She works in the field of Computational Linguistics since 1972, first as Researcher at the Department of Linguistics of Pisa University, then as Director of Research at ILC, Istituto di Linguistica Computazionale of CNR, Pisa.

Since August 2003 she is the Director of the Istituto di Linguistica Computazionale of the Italian

Research Council (ILC-CNR), Pisa, Italy.

She has co-ordinated many international, European and national projects and strategic initiatives, mostly in the fields of *Language Resources and Standardisation*. She has more than 350 publications. In addition to other editorial activities, she is Director of the Journal *Linguistica Computazionale*, IEPI, Pisa – Roma, and Co-editor with Nancy Ide of the new International Journal *Language Resources and Evaluation*, Springer. Conference chair of LREC2004, LREC2006, COLING/ACL2006, Italian TAL2006. Main fields of interest are: Human Language Technology; computational lexicology and lexicography; language resources; corpus linguistics; standardisation and evaluation of language resources; lexical semantics and semantic annotation; collocations and multiwords; derivational morphology; knowledge acquisition from multiple (lexical and textual) sources, integration and representation; validation of language resources.

Prof. Calzolari is, among others, member and general secretary of ICCL, Vice-president of the ELRA Board, President of the PAROLE Association, founding member of the Italian Forum for HLT at the Ministry of Communications, convenor of WG4 of ISO TC37 SC4, member of the Advisory Committee for the 21st Century COE (Center of Excellence) Program of Tokyo Institute of Technology, member of IULA (Barcelona) Advisory Committee, member of SENSEVAL Advisory Committee, member of many other International Committees and Advisory Boards. Invited speaker, member of program committee or organiser for quite numerous international scientific conferences, workshops, etc.