WP5 Task 4 State-of-the-Art Natural Language Processing

Sections:

Morphological Analysis -- Sebillot/IRISA
Part-of-speech tagging -- Debili/CNRS
Automatic Dictionary Extraction for Bilingual Text -- Grefenstette/CEA
Language Resources for Less Studied Languages -- Grefenstette/CEA
Automatic Ontology Creation/Extension -- Sebillot/IRISA
Monolingual Information Retrieval Kotropoulos/AIIA
Cross Language Information Retrieval -- Rauber/TU Wien-IFS
Text Classification -- Novovicova/UTIA

Morphological Analysis

Sebillot/IRISA

1- Description

Morphological analysis is concerned with the inflectional, derivational, and compounding processes in word formation. It corresponds to the segmentation of a given word into the various smallest meaning units (morphemes) which form it, *e.g.* its stem and affixes. A full morphological analysis can also give morphosyntactic information about the word-form and its stem, *e.g.* possible part-of-speech (PoS) and/or inflectional properties (gender, number, case, person, tense, etc.), etc. Morphological analysis is a key-point for a lot of NLP applications. It is often realized together with PoS tagging during the pretreatment phase of a corpus, and permits to recognize the presence of a same word or concept in spite of different morphological variants.

2- Current approaches

A first, simple approach of morphological analysis consists in using available electronic lexical databases that associate word-forms and lemmas, together with inflectional and/or derivational information. This last information is often provided as a code or model to which operations to produce all possible inflected forms are attached. For example, the MULTEXT project (Armstrong 96) has provided lexical lists of lemmas and inflected word-forms for four languages of the European Community: French, Italian, Spanish and German. The word-form list contains word-forms, lemmas and a linguistic description, which encodes features that have been considered relevant for several languages and are based on EAGLES (Expert Advisory Group on Language Engineering Standards) recommendations for computational lexicons. The word-form dictionary for French lists 300,000 forms including proper nouns and compounds. Each character of the linguistic description specifies a value of an attribute. For a verb, there are 7 attributes: PoS, type, mood or verbal form, tense, person, number, gender; for a noun, 5 attributes: PoS, type, gender, number, and case; and for an adjective, there are 6 attributes: PoS, type, degree, gender, number, and case. Compounds receive the same linguistic description as simple words. CELEX (Burnage 90) is a large multilingual database that includes extensive lexicons of English, Dutch, and German. For each language, several types of lexicons are available: lemma, word-form, abbreviation and corpus type. In the lemma lexicon, each entry represents a full inflectional paradigm. In the word-form lexicon, entries deal specifically with one flexion. The corpus type lexicon contains strings extracted from various contemporary texts and is a representative list of real-life words, distinguished on the basis of their spelling, with detailed information about their frequency. Thus, inflectional information is present in the word-form lexicon and derivational information in the lemma lexicon.
This approach has however difficulties to deal with the quasi infinite possibilities of the derivational process, and offers no (efficient) way to analyze words not present in the database.

A second view concerns morphological systems. The simplest ones are stemmers, among which the best known stemming algorithms are those developed by Lovins (68) and Porter (80). The common point between the different stemmers, which treats inflectional and derivational affixes identically, is to proceed in two steps: the first is the de-suffixing step which consists in withdrawing predefined endings from words; the second is the recoding

phase which adds predefined endings to the previously obtained roots. Those two phases can be performed successively, as in Lovins's stemmer, or simultaneously, as in Porter's. For example, Porter's stemmer relies on a set of transformational rules such as *–ational ® -ate* which transforms a word such as *relational* into *relate.* Words are coded in pseudo-syllables so as to avoid applying the stemming procedure on too short words. Porter's stemmer only reduces suffixes; prefixes or compounds are not simplified. Though a language such as English can be morphologically analyzed with stemming methods, this is not the case for highly inflectional languages which require more sophisticated techniques. Koskenniemi (1983) has proposed a model of two-level morphology with encompasses both morphotactics, the ordered decomposition of a word into morphemes, and morphophonemics, the alternate forms of morphemes according to the phonological context. For example, the word *specifies* is analyzed as the stem *specify* and the suffix *–s*. The addition of the suffix *–s* transforms the final *y* of *specify* into *ie*; thus *specify* and *specifie* are allomorphs. In this model a word is thus represented as a correspondence between its lexical level form and its surface level form. The two-level morphology model has been implemented with a special kind of finite state automata, finite state transducers, which allows the encoding of a correspondence between two letters, one belonging to the surface form and the other belonging to the lexical form. For example, the word *specifies* receives the following representation, where + is a morpheme boundary symbol: lexical form: *s p e c i f y + s*, and surface form: *s p e c i f i e s*. The first implementation of two-level morphology led to the system KIMMO (Karttunen83), which has two analytical components: the rule component and the lexicon. The rule component consists of two-level rules that account for regular orthographical alternations. The lexicon lists all morphemes (stems and affixes) in their lexical form and their morphotactic constraints. KIMMO has two processing functions: the Generator and the Recognizer. The Generator accepts as input a lexical form such as *specify+s* and returns the surface form *specifies*. The Recognizer accepts as input a surface form such as *specifies* and returns a form divided into morphemes, such as *specify+s*, plus a gloss string *verb+present+3sg*. In 1993, PC-KIMMO (version that runs on a variety of systems) was enhanced by a third analytical component, a word grammar, that provides parse trees and feature structures. Much state-of-the-art systems for morphological analysis of word-forms are usually based on those two-level finite-state transducers.

If inflectional morphology is currently well treated, both derivational and compounding processes still present difficulties that motivate most recent researches in morphological analysis[1]. Concerning derivation, the problem of controlling the possible and impossible derivations, and of interpreting the semantic value of these transformations are for example deeply studied in the MorTAL project and the GéDérif system. And in order to palliate difficulties of the two-level systems to deal with concatenative compound morphology, and the cost and complexity of hand-crafting two-level rules despite the availability of sophisticated development tools, emphasis is put on machine learning techniques, and on the study of agglutinative languages and extensions of the two-level model to correctly deal with their problematics. Memory-based approach has for example been chosen as an alternative supervised machine learning technique for morphological analysis. And the possibilities of SVM (support vector machines) have been studied on Arabic or Japanese morphology. In a different purpose, let us also mention Theron and Cloete's work that uses machine learning methods to automatically acquire two-level rules. Concerning the second point, the study of agglutinative language morphology has led to extensions of the two-level model, particularly to be able to treat non sequential morphotactic constraints (long distance dependencies). A

[1] References concerning these recent works are given in the bibliography section.

three-level model has been proposed, where a feature-based third level intends to encode the morphological properties of the input word; the need to impose a hierarchical structure upon sequences of morphemes and to build complex constructions from them has for example led to a combination of a two-level and a unification-based formalisms to treat morphology of the Basque language; finite-state morphological transducers and complex descriptions based on typed feature structures have also been integrated.

References

Armstrong S. MULTEXT: Multilingual Text Tools and Corpora, Lexikon und Text, Tübingen: Niemeyer, H. Feldweg and W. Hinrichs eds., pages 107-119, 1996

Burnage G. CELEX: A Guide for Users, Center for Lexical Information, University of Nijmegen, 1990

Lovins J.B. Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, pages 22-31, 1968

Porter M.F. An Algorithm for Suffix stripping, Program, Vol. 14, pages 130-137, 1980

Koskenniemi K. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production, PhD Thesis, University of Helsinki, 1983

Karttunen L. KIMMO: A General Morphological Processor, Linguistic Forum, Vol. 22, pages 163-186, 1983

3- Bibliography

Aduriz. I., Agirre E., Aldezabai I., Alegria I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., and Urkia M. A Word-Grammar Based Morphological Analyzer for Agglutinative Languages, Proceedings of Coling 2000, 18th International Conference on Computational Linguistics, Saarbrücken, Germany, 2000

Daille B., Fabre C., and Sébillot P. Applications of Computational Morphology, In Many Morphologies, pages 210-234, Cascadilla Press, Somerville, P. Boucher ed., 2002

Diab M., Hacioglu K., and Jurafsky D. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of HLT-NAACL 2004, Boston, USA, 2004

Namer F., Dal G. GéDériF: Automatic Generation and Analysis of Morphologically Constructed Lexical Resources, Proceedings of LREC 2000, second International Conference on Language Resources and Evaluation, Athens, Greece, 2000

Nakagawa T., Kudo T., and Matsumoto Y. Morphological Analysis Using Support Vector Machines and Proposal of Revision Learning, IPSJ Journal, Vol.44, No 05, 2003

Theron P., and Cloete I, Automatic Acquisition of Two-Level Morphological Rules, Proceedings of ANLP97, Applied Natural Language Processing, Washington, DC, USA, 1997

Van den Bosch A., and Daelemans W. Memory-based Morphological Analysis. Proceedings of ACL'99, 37th Annual Meeting of the Association for Computational Linguistics, Maryland, USA, 1999.

4- URLs

http://chasen.aist-nara.ac.jp/
http://ilk.uvt.nl/mblem/
http://lit.dfki.uni-sb.de:8000/annotation/index.html

http://www.cs.technion.ac.il/~erelsgl/bxi/hmntx/teud_tokna.html#English
http://www.issco.unige.ch/projects/MULTEXT.html
http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman-form-e.html

http://www.lingsoft.fi/demos.html
http://www.sil.org/pckimmo/
http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.html
http://www.xrce.xerox.com/

Part-of-Speech Tagging

Debili/CNRS

1- Description

Words are frequently ambiguous: in French, *ferme* is potentially *noun, verb, adjective* or *adverb*. In English, *man* is potentially a *noun* or a *verb*. In Arabic, ??????? (*ktb*) could be a *verb* or a *noun*. Part-of-speech (POS) tagging is the operation of assigning the proper grammatical category (*noun, verb, adjective,* etc.) to each word of a given text according to the context in which it appears. The following sketch demonstrates the combinatory complexity of the problem:

w_1	w_2	w_3	...	w_N
t_{11}	t_{21}	t_{31}	...	t_{N1}
t_{12}	t_{22}	t_{32}	...	t_{N2}
…	…	…	...	…
t_{1p}	t_{2q}	t_{3r}	...	t_{Ns}

In the above figure, w_i are the words in the sentence or text, and t_{ij} are the respective different grammatical tags each word may have out of context. If the average number of potential grammatical tags of a given word is 2, then a 25 word sentence could be tagged in 2^{25} different ways. How can the computer choose among these millions of possible grammatical configurations, the single (or possible few) correct assignment of tag to word?

Actually, this sketch simplifies the problem too much, presupposing that the word segmentation has already been performed. In reality, the distinction of the w_i for example with sequences of words such as "*as much as*" or, in French "*bien que*" or many of the agglutinated words of Arabic or Turkish, may be undecidable until tagging begins.

Since the 1960s, POS tagging has been object of numerous studies in computational linguistics, dealing with how grammatical tags should be defined, and, once defined, with accurate and efficient algorithms for choosing POS tags for the words in a text. POS Tagging has been considered as a useful preliminary step for lemmatization and syntactic analysis, exploited in many natural language processing applications (machine translation, information retrieval, speech recognition. But decades of research producing results that remain difficult to evaluate show that POS tagging covers many open problems.

2- Current approaches

Problems arise on the linguistic level when we try to define a set of tags for a language, and then the subsequent rules for choosing among the tags. Computationally, we must resolve problems of exponential time and space. The main intuitions for solving the computational problems are found in most current approaches:

- Implement rules using limited context to define sequences of permissible parts-of-speech,
- use training techniques to build these succession rules,
- use frequency as a basis for deciding among competing rules,
- use ad hoc rules to treat remaining ambiguity

Tags: many authors admit that this is the Achilles' heel of the POS tagging. One frequently sees the same research team give various sets of grammatical tags for the same language, the definitions of which are not clearly stabilized. This indecision about the proper tag set is aggravated when sets from different research teams are compared. The number of tags defined can number from dozens to hundreds, with different ideas of what each tag covers, making comparative evaluation between published results difficult to perform. In practice, we see a slow, cumulative and consensual work which is being performed by teams who are compelled to put up with lists with undefined contours and who try to update them progressively.

Rules: involve exploiting the context around an ambiguous word. Context is generally considered as a window of words (and their possibly disambiguated tags) around the ambiguous word, though it might include other discriminating criteria which happen to be discriminating such as the presence of a capital letter at the beginning of one word, the presence of particular morphemes, etc. These rules are handmade or automatically built from training texts, that are usually pretagged.

The most successful approaches are those which extract automatically their rules from large amounts of pre tagged text. The number of rules could easily become very large, so the amount of context that can used by a rule is often constrained (usually by a window size based on the number of words before or after an ambiguous word). Considering that many rules could apply simultaneously and therefore numerous different tags could follow from these applications, weights, scores, or probabilities are often attached to these rules to order or to determine their application.

Algorithms: The main problem comes from the exponential number of possible tag sequences. Computations often use a variant of the Viterbi algorithm, based on dynamic programming, which reduces the problem from N^L complexity to LN^2 complexity (where N is the number of tags and L the length of the sentence).

Unknown words: Tags attached to words come frequently from dictionaries, or from morphological analysis. For the non identified words, tags come from pre established lists, or from dynamically built lists according to properties of neighbour words.

In practice, morphological analysis and POS tagging tends to be more and more intertwined, with tagging being perceived as a complementary phase allowing the morphogrammatical description (words, lemmas, tags) of input text to be essentially correct.

Evaluation: as explained above, evaluation of POS taggers is difficult, although common experimental bases are being built up due to increased availability of shared linguistic resources (dictionaries, corpora, pre tagged corpora, etc.) which make a comparative evaluation possible.

Published accuracies are mainly related to the rate of correctly tagged words, and rarely to the rate of correctly tagged sentences, with results often in the 96% - 97% range for words. Considering entire sentences, accuracy rates seem not to exceed 60%.

3- Bibliography

Adda, G., Mariani, J., Paroubek, P., Rajman, M., & Lecomte, J. (1999). "L'action GRACE d'évaluation de l'assignation des parties du discours pour le français". Langues, 2(1), 119-129.

Eric Brill. A simple rule-based part-of-speech tagger. In Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing, pages 152-155, Trento, IT, 1992.

E. Charniak, G. Carroll, J. Adcock, A. Cassandra, Y. Gotoh, J. Katz, M. Littman, and J. McCann. Taggers for parsers. Artificial Intelligence, 85(1-2):45-57, 1996.

Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. Equations for part-of-speech tagging. In National Conference on Artificial Intelligence, pages 784-789, 1993.

Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the Second Conference on Applied Natural Language Processing, pages 136-143, Austin, Texas, 1988.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, 1992.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. MBT: A memory based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, Proceedings of the Fourth Workshop on Very Large Corpora, pages 14-27, 1996.

Steven J. DeRose, 1988. "Grammatical Category Disambiguation by Statistical Optimization". *Computational Linguistics* Volume 14, Number 1, 31-39.

Hans van Halteren, editor. 1999. *Syntactic wordclass tagging*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

H. van Halteren, J. Zavrel, and W. Daelemans. Improving accuracy in word class tagging through the combination of machine learning systems. Computational Linguistics, 27(2):199-230, 2001.

Klein, S. & R. F. Simmons. 1963. A Computational Approach to Grammatical Coding of English Words. *Journal of the Association for Computing Machinery*, 10: 334-347.

Merialdo, Bernard. 1994. "Tagging English text with a probabilistic model". *Computational Linguistics* 20.2: 155-172.

Emmanuel Roche and Yves Schabes. Deterministic part-of-speech tagging with finite-state transducers. Computational Linguistics, 21(2):227-253, 1995.

H. Schütze and Y. Singer. Part-of-speech tagging using a variable memory markov model. In Proceedings of the 32nd Annual Meeting of the ACL, 1994.

Pasi Tapanainen and Atro Voutilainen. Tagging accurately - don't guess if you know. In Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP'94), pages 47-52, Stuttgart, Germany, 1994.

Kristina Toutanova, Dan Klein, Christopher D. Manning, Yoram Singer. (2003). "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". HLT-NAACL, 2003.

André Valli, Jean Véronis. 1999. "Etiquetage grammatical des corpus de parole: problèmes et perspectives", *Revue Française de Linguistique Appliquée*.

4- URLs

http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html
http://www.xrce.xerox.com/competencies/content-analysis/demos/english
http://www.lingsoft.fi/cgi-bin/engcg
http://ilk.kub.nl/~zavrel/tagtest.html

Automatic Dictionary Extraction for Bilingual Text

Grefenstette/CEA

1- Description

A bilingual dictionary provides a, possibly ranked, set of translations for a given word or term. Bilingual dictionaries are needed for human translation of technical texts, machine translation, and cross-language information retrieval. Constructing such dictionaries for new language pairs, or keeping them up-to-date for new domains is an expensive process. As more and more translated pairs of documents (called *bitexts*) become available electronically, research has been advancing on exploiting these texts to automatically create bilingual dictionaries for new domains or for new language pairs.

2- Current approaches

Since Champollion used the Rosetta stone to decipher hieroglyphics in 1822, there has been interest in using parallel texts to develop bilingual dictionaries. The premise of automatic bilingual dictionary extraction is that, given enough *bitext*, the statistics of term co-occurrence allows us to discover which source words are translated by which target language words. The first step in this procedure is to align text segments between the two versions of the *bitext*. Most systems use the sentence as the text segment. The types of sentence alignments that are usually considered are 1-1 (one sentence translated by another), 1-2 (one sentence translated by two sentences), 2-1, 2-2, and 0-1 (a new sentence added) and 1-0 (a sentence deleted). Sentence alignment techniques can achieve 96% success rates.

Kay and Röscheisen (1989) proposed one of the first methods for aligning sentences using an iterative calculation of word co-occurrence distributions to identify anchors between segments. Gale and Church (1991) described a method based only on comparing sentence lengths. Others (Brown et al., 1991, Simard et al., 1992) proposed using cognates, i.e., strings that are similar in both languages such as numbers, proper names, and punctuation, or even small bilingual lexicons (Catizone, et al., 1989, Debili and Sammouda, 1992) to improve sentence alignment. The Gale and Church method is the most widely used since the source code was published in the 1993 version of the paper.

Once the sentence segments have been aligned, statistics on co-occurrence between words are calculated over the entire collection of text. These statistics are used to predict translations (Hiemstra, 1998). Some have extended this work to phrases, for example Hull (1998) used a part-of-speech tagger to chunk each side of the *bitext*, and then collected statistics on these chucks rather than individual words. By the end of the 1990s, text and word alignment was becoming a well studied problem, thanks to the ARCADE competition (Veronis, 1999) and to experimentation and development during the John Hopkins summer school in 1999 at the Center for Language and Speech Processing. The latter course produced GIZA++ (Och et al, 1999), a word and term alignment freeware. GIZA++ was widely used in the recent cross-language campaigns of CLEF and TIDES to generate automatic translation dictionaries for rare pairs of languages, for example Hindi-English (Larkey et al. 2003).

Much recent work in creating automatic bilingual dictionaries involves building aligned corpora from the Web. Collections are either found and treated manually, or they can be

found automatically using a program such as STRAND (Resnick and Smith, 2003). STRAND exploits the fact that many web pages containing pointers to versions of the page in a different languages, for example, using language names in the link, and the fact that many websites with multilingual web pages are built with parallel file structures that can found in the URL names. The mark-up structure and the lengths of suspected parallel pages are compared, and only likely matches are fed into sentence alignment and word alignment programs. This Web-based approach offers great promise for automatic bilingual dictionary creation as the Web continues to grow and expand to rarer languages and continued use of vernacular languages with English on the same web sites.
A recent workshop called The Amazing Utility of Parallel and Comparable Corpora http://132.167.34.85/~lic2m/bibliographie/conferences/lrec-2004/ws/ws11.pdf presents the latest results on using parallel corpora to extract bilingual lexicons.

3- Bibliography

Brown, P, Lai, J., and Mercer, R. (1991), `Aligning Sentences in Parallel Corpora', in Proceedings of 29th Annual Meeting of the Association for Computational Linguistics, (Morristown NJ), 169-176.

Catizone, R., G. Russel, and S. Warwick. (1989) Deriving translation data from bilingual texts. In Uri Zernik, editor, Proc. of the First Int. Lexical Acquisition Workshop, Detroit

Debili, F. & Sammouda, E. (1992). Appariement des phrases de textes bilingues franais-anglais et franais-arabe, Proc. 15th International Conference on Computational Linguistics (Coling 92) Nantes, France

Gale, W. and K. W. Church (1991). A program for aligning sentences in bilingual corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.. Also published in Computational Linguistics, 19 (1): 75--102, 1993

Harris, B. (1988). "Bi-text, a new concept in translation theory." Language Monthly 54 pp. 8-10

Hiemstra, Djoerd (1998) Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus', In: Peter-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds.) Proceedings of the eightth CLIN meeting, pp. 41-58

Hull, D. A. (1998). A Practical Approach to Terminology Alignment. Proceedings of Computerm '98 (First Workshop on Computational Terminology). Montreal, pp 1-7.

Kay, M. and M. Röscheisen (1993). "Text-Translation Alignment." Computational Linguistics 19(1) pp. 121-142

Melamed, D. I. (1996). Automatic Construction of Clean Broad-Coverage Translation Lexicons. Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96). Montreal

Larkey, Leah S., Margaret E. Connell, Nasreen Abdul Jaleel (2003) Hindi CLIR in thirty days. ACM Trans. Asian Lang. Inf. Process. 2(2) pp. 130-142

Och, Franz Josef, Christoph Tillmann, and Hermann Ney. (1999) Improved alignment models for statistical machine translation. In Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20-28, University of Maryland, College Park, MD, June

Resnik, Philip and Noah A. Smith (2003). "The Web as a Parallel Corpus." Computational Linguistics 29(3):349-380, September (special issue on the Web as a corpus).

Simard, Michel, George F. Foster, and Pierre Isabelle. (1992) Using cognates to align sentences in bilingual corpora. In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI), pages 67-81, Montreal

Véronis, J. and P. Langlais (1999). Evaluation of parallel text alignment system - The ARCADE project. Parallel Text Processing. J. Véronis. Berlin, Kluwer.

4- URLs

http://www.up.univ-mrs.fr/veronis/arcade/ARCADE
First stage included a competition among 6 systems for sentence alignment. ARCADE II to start soon. Some studies on evaluation. Evaluation measures include standard precision, recall and F-measure.

Software:

- http://wwwhome.cs.utwente.nl/~irgroup/align/download.html Twente (Hiemstra 1998)
- http://www.clsp.jhu.edu/ws99/projects/mt EGYPT Includes an word-alignment visualization tool Cairo. Languages: Arabic, French, Czech,Timorese.
- http://www.d.umn.edu/~tpederse/parallel.html ALPACO, KVEC
- http://numerus.ling.uu.se/~corpora/plug/pwa/ PLUG. Word Alignment tool available free of charge. (binaries) Demo available for Swedish-English corpora. Languages: Swedish
- http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html GIZA ++ , Improvement over GIZA; includes an implementation of the models in F. Och, H. Ney. "Improved Statistical Alignment Models".

Bitext resources:

- http://www.isi.edu/natural-language/download/hansard/ Hansards, Canadian parliament
- http://www.umiacs.umd.edu/~resnik/strand/STRAND databases
- http://europa.eu.int European Union parallel texts
- http://nl.ijs.si/ME/CD/mte-home.html MULTEXT-EAST. Sentence alignment of Orwell's 1984 in languages from East Europe.Languages: Bulgarian, Czech, Estonian, Hungarian, Lithuanian, Latvian, Romanian, Russian, Serbo-Croatian, Slovene.
- Lists of bilingual corpora:
    - http://stp.ling.uu.se/~corpora/
    - http://www.corpus-linguistics.de/corpora/corp_parallel.html

- http://www.nilc.icmc.usp.br/nilc/tools/parallelcorpora.htm
- http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/types/parallel.html

Language Resources for Less Studied Languages

Grefenstette/CEA

1- Description

For most natural language processing applications, such as information extraction, retrieval and filtering, a number of language-specific resources and tools are needed. These resources range from text tokenizers and segmenters, lexicons, part-of-speech taggers and syntactic analysers. The purpose of these resources and tools is to be able to reduce textual variations to canonical forms that can be exploited by a computer. At the current time, most research in natural language processing has been devoted to English for historical and economic reasons. In the last ten years as the Internet has made the computer an essential part of many local economies, more research has been devoted to local, less studied languages, but resources for these languages are still lacking.

2- Current approaches

Natural language processing received great impetus for the English language by the creation of two large corpora: the Brown Corpus in the early 1960s and the British National Corpus in the late 1980s. During the 1990s, many language resources for treating English were developed, thanks in part to these corpora. In the late 1990s, resources became more abundant for other western European languages, but many languages for which a written tradition exists still did not possess adequate lexical resources for advanced language applications. With the appearance of the Internet, there arose a fear that English would become the dominant language to the exclusion of all others, though other languages seem to be growing at a faster pace than English (Grefenstette & Nioche, 2000). UNESCO edits a list of languages that are in danger of disappearing (Wurm, 2001). This situation has motivated a number of researchers to attack the problem of creating resources for less-studied languages.

One of the first tasks for creating resources for a rare language is to build a corpus for that language. Ghani et al. (2001) present a technique that, from one example document in a given language, crawls the Web to retrieve other, different pages in the same language. This supposes that the language is fairly well represented on the Web, a statement that holds for probably about 80 languages. Work on building initial lexicons for an unknown language, uniquely from corpora, has been undertaken by researchers such as Goldsmith (2001), who finds the statistically most likely combinations of stems and suffixes to cover a list of words. For languages written without spaces, such as Chinese and Japanese, building a lexicon involves finding the best segmentation of an input text, usually in an iterative process (Wu & Fung, 1994). Once a small lexicon has been built, it is possible to use induce paradigms from the lexicon, and to use Web frequencies to predict and validate additional words and their normalizations to be added to the lexicon (Grefenstette, et al.).

See the section in this report on "Automatic Dictionary Extraction for Bilingual Text" for a description of techniques for building translation lexicons for less studied languages. Here is one example of such work (Weerasinghe, 2002). A recent workshop at LREC'2004 was organized by called *First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilatio*n contains many papers on using bilingual corpora for creating lexicons (bilingual) involving rare languages.

A US-government sponsored program called TIDES (Translingual Information Detection and Extraction, and Summarization) (http://www.darpa.mil/ipto/programs/**tide**s/) has held "Surprise Language Competitions" in which research groups had 30 days to prepare translation resources for an unknown (surprise) language. A presentation of the 2003 competition can be found in the URLs listed below.

There has also been a largely manual effort to extend the thesaurus-like resource called WordNet to other languages (Vossen at al, 2001), such as Basque (Agirre et al. 2002).

3- Bibliography

Agirre E., Ansa O., Arregi X., Arriola J.M., Diaz de Ilarraza A., Pociello E. and Uria L. *Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis* Proceedings of the first International WordNet Conference in Mysore, India, 21-25 January 2002

Ghani, R. and R. Jones, and D. Mladenic. Mining the web to create minority language corpora. In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM), 2001.

Goldsmith, J. (2001) Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics 27, 2, 153-198

Grefenstette, G. and J. Nioche. *Estimation of English and non-English language use on the WWW*. Proc. RIAO 2000, Content-Based Multimedia Information Access, pages 237-- 246, 2000.

Gregory Grefenstette, Yan Qu, and D. A. Evans, Expanding lexicons by inducing paradigms and validating attested forms. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)* , Las Palmas, Canary Islands, Spain, 2002

Vossen P., Bloksma L., Climent S., Marti M.A., Taule M., Gonzalo J., Chugur I., Verdejo F., Escudero G., Rigau G., Rodriguez H., Alonge A., Bertagna F., Marinelli R., Roventini A., Tarasi L., Peters W. (2001). *Final Wordnets for Dutch, Spanish, Italian and English, EuroWordNet (LE2-4003)* Deliverable D032/D033, University of Amsterdam (The Netherlands). http://www.illc.uva.nl/EuroWordNet/

Weerasinghe, Ruvan (2002). Bootstrapping the Lexicon Building Process for Machine Translation between `New' Languages. Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA. Pp. 177-186.

Wu, D. and Fung, P. (1994). Improving Chinese tokenization with linguistic filters on statistical lexicon acquisition. In Proceedings of the Fourth Conference on Applied Natural Language Processing, pages 180-181.

Wurm, Stephen A.(ed.) 2001. Atlas of the World's Languages in Danger of Disappearing. Paris: UNESCO publishing. Further information available at: http://upo.unesco.org/bookdetails.asp?id=1352

4- URLs

http://isl.ntf.uni-lj.si/SALTMIL/ The ISCA (International Speech Communication Association) Special Interest Group on Speech and Language Technology for Minority Languages.

http://linguistica.uchicago.edu/ Linguistica group at the University of Chicago, unsupervised learning of natural language structure. See also John Goldsmith's page http://humanities.uchicago.edu/faculty/goldsmith/StatNLP/

http://www.glue.umd.edu/~oard/papers/mitre.ppt Douglas Oard's presentation on the 2003 TIDES surprise language task

Automatic Ontology Creation/Extension

Sebillot/IRISA

1- Description

Following Gruber's definition (1994), ontologies are formal, explicit specifications of shared conceptualizations, representing concepts and their relations that are relevant for a given domain of discourse. Ontologies are for example used as shared references between remote applications and agents in the Semantic Web vision. There are in fact a lot of levels of description among ontologies. Formal ontologies may be opposed to terminological ones (Sowa 2000). In the former, categories are distinguished by axioms and definitions stated in logic or other computer-oriented languages, which support complex inferences and computations. In the latter, from which WordNet (Fellbaum 1998) is a prototype, categories are partially specified by relations such as super/subtype or part-whole, which determine the relative positions of the concepts but do not completely define them. Indeed terminological ontology definition goes from simple lexicons or controlled vocabulary to thesauri, taxonomies with hierarchical relations between terms, or ontologies with named concepts. Some editing tools or environments have been developed in order to ease ontology construction (Kaon, OntoEdit, Protégé, WebOde, etc.). However there are still based on a large part of manual work. Automation of ontology construction can be reached by a combined use of NLP and machine learning techniques applied to texts of the concerned domain. Those techniques may be used to directly build an ontology from a corpus or to update and refine an existing one to fully adapt it to a domain. The presentation hereafter is thus focused on current approaches of ontology creation and/or extension from texts.

Fellbaum C., ed., WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998
Gruber T. Towards Principles for the Design of Ontologies Used for Knowledge Sharing., International Journal of Human and Computer Studies, Vol. 43, No 5/6, pages 907-928, 1994
Sowa J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000

2- Current approaches

A typical way to process in this field of ontology learning from text is to extract simple and/or complex terms of the domain from a textual corpus, and to cluster them into groups, trying to identify a taxonomy of potential classes. Numerous techniques and systems to catch domain-specific terms have been proposed, together with works and methods concerned with semantic class creation, but also with the detection of different kinds of relations (vertical, horizontal, or transversal ones) between terms and conceptual classes of terms. Indeed, techniques to grasp both terms and relations are quite similar. Among this huge domain, it is possible to oppose numerical *versus* symbolic techniques of acquisition from corpora. Numerical approach of (term or relation) acquisition exploits the frequential aspect of data, and uses statistical techniques; on the other hand, symbolic approach exploits the structural aspect of data, and uses structural or symbolic information.

Within the numerical approach, complex terms or (syntagmatic) relations between terms can be acquired by studying word co-occurrences within a text window, and evaluating the strength of the association with the help of a statistical score (association coefficient) that detect words appearing together in a statistically significant way. Following Harris's linguistic principles (1989), numerical distributional analysis methods respect a 3-step approach: extraction of the co-occurrents of one word (within a text window or a syntactic context), evaluation of proximity/distance between two terms, based on their shared or not shared co-

occurrents (various measures are defined), clustering into classes (following different data analysis or graph techniques for example).

The symbolic approach of acquisition groups in fact two visions: symbolic linguistic approach, and symbolic machine learning (ML) approach. In the first one, operational definitions of the elements to acquire are established manually by linguists, usually in the form of morpho-lexical patterns that carry the complex terms or relations that are studied, or by a list of linguistic clues. However when such patterns or clues are unknown, but examples of elements respecting the target relation or form are known, symbolic ML can be used to automatically extract patterns from the descriptions of those examples. The technique is based on a 5-step methodology initiated by Hearst (92): 1- select one target relation R; 2- gather a list of pairs following relation R; 3- find the sentences that contain those pairs; keep their lexical and syntactic contexts; 4- detect common points between those contexts; suppose that they form a pattern for R; 5- apply the patterns to get new pairs and go back to 3. Symbolic ML (inductive logic programming, grammatical inference, etc.) offers a framework to automate step 4, and automatically produce the unknown morpho-lexical patterns that carry the target kind of terms or relation.

Both approaches present advantages and drawbacks. Numerical approach is portable, automatic but produces non interpretable results; the detection is realized at the corpus level: one occurrence (kept or not) cannot thus be explained; and rare cases are problematic. Symbolic approach needs *a priori* knowledge (patterns, examples), but produces interpretable results; detection is done at the occurrence level, and rare cases can be treated. In order to take advantages from both, a combination of one method from each family is often used to obtain an interesting mixed solution.

Among the numerous complex-term extractors, Termino (David and Plante 1991) is for example based on a symbolic linguistic approach; Church and Hanks's work (1989) is based on a statistical co-occurrence technique; and Acabit (Daille 1996) on a mixed one. Grefenstette (1994) presents a description of statistical ways to extract relations and information from texts, and Claveau *et al.* (2003) a symbolic ML system.

The use of various techniques in order to create more "exhaustive" structured ontologies is a really up-to-date domain. Syntex/Upery (Bourigault 2002) is, for example, a tool developed as an aid for ontology engineers and experts to build a complete ontology by only mining texts of the concerned domain. And OntoLearn (Navigli and Velardi 2004) is an automatic tool to refine and adapt an existing ontology (WordNet) to a precise domain.

Evaluating the obtained ontologies is quite a deep and open problem. Various conferences and workshops are dedicated to evaluation (about this problem, see for example Deliverable 1.4 at http://www.ontoweb.org/deliverable.htm on the OntoWeb thematic network site). Both extracting tools and resulting ontologies have to be evaluated. For the former, some evaluation campaigns are organized (*e.g.* the Cesart part of Evalda campaign about terminology extraction http://www.technolangue.net/article58.html). For the latter, a few platforms and criteria are provided (see for example the unpublished report by Angele and Sure from Ontoprise http://www.ontoprise.de/documents/effort-ekaw.pdf). The ending conclusion still remains in the usability of the ontology by the application it is dedicated to.

Church K.W., and Hanks P. Word Association Norms, Mutual Information, and Lexicography, Proceeding of ACL'89, 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 1989

Claveau V., Sébillot P., Fabre C., and Bouillon P. Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming, Journal of Machine Learning Research, Special Issue on Inductive Logic Programming, Vol. 4, pages 493-525, 2003

Daille B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, in The Balancing Act: Combining Symbolic and Statistical Approaches to Language, P. Resnik and J. Klavans eds, MIT Press, pages 49-66, 1996

David S., and Plante P. Le progiciel TERMINO : de la nécessité d'une analyse morpho-syntaxique pour le dépouillement terminologique de textes. Proceedings of Colloque sur les industries de la langue, Québec, Canada, 1991
Grefenstette G. Explorations in Automatic Thesaurus Discovery, Dordrecht: Kluwer Academic Publishers, 1994
Harris Z.., Gottfried M., Ryckman T., Mattick P. (Jr), Daladier A., Harris T.N., and Harris S. The Form of Information in Science, Analysis of Immunology Sublanguage, Kluwer Academic Publisher, Dordrecht, 1989
Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of Coling'92, 14th International Conference on Computational Linguistics, Nantes, France, 1992
Bourigault D. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Proceedings of TALN 2002, 9e conférence annuelle sur le Traitement Automatique des Langues, Nancy, France, 2002
Navigli R., and Velardi P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, Computational Linguistics, Vol. 30, No 2, pages 151-179, 2004

3- Bibliography

Alfonseca E., and Manandhar S. Improving an Ontology Refinement Method with Hyponymy Patterns. Proceedings of Language Resources and Evaluation, LREC-2002, Las Palmas, Spain, 2002
Buitelaar P., Olejnik D., Sintek M. A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. Proceedings of the 1rst European Semantic Web Symposium ESWS-2004, Heraklion, Greece, 2004
Claveau V., and Sébillot P. From Efficiency to Portability: Acquisition of Semantic Relations by Semi-Supervised Machine Learning. Proceedings of Coling 2004, 20th International Conference on Computational Linguistics, Geneva, Switzerland, 2004
Faure D., and Nédellec C. ASIUM: Learning Subcategorization Frames and Restrictions of Selection. Proceedings of the Text Mining workshop, 10th European Conference on Machine Learning (ECML 98), Kodratoff Y. ed., Chemnitz, Germany, 1998
Gomez-Perez A., and Manzano-Macho D. A Survey of Ontology Learning Methods and Techniques. Deliverable 1.5, OntoWeb Project, 2003
Maedche, A., and Staab, S. Semi-automatic Engineering of Ontologies from Text, Proceedings of Seke2000, International Conference on Software Engineering and Knowledge Engineering, Chicago, USA, 2000
Navigli R., and Velardi P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, Computational Linguistics, Vol. 30, No 2, pages 151-179, 2004
Le Moigno S., Charlet J., Bourigault D., Degoulet P, and Jaulent M.-C. Terminology Extraction from Text to Build an Ontology in Surgical Intensive Care, Proceedings of AMIA Annual Symposium 2002, San Antonio, USA, 2002

4- URLs

http://kaon.semanticweb.org/Members/rvo/Module.2002-08-22.4934
http://ontoweb.aifb.uni-karlsruhe.de/
http://www.cs.utexas.edu/users/mfkb/related.html
http://www.dfki.de/~paulb/esws04.pdf
http://www.dsi.uniroma1.it/~velardi/IEEE_C.pdf
http://www.lri.fr/%7Efaure/Demonstration.UK/Presentation_Demo.html
http://www.univ-tlse2.fr/erss/membres/bourigault/

Monolingual Information Retrieval

Kotropoulos/AIIA

1- Description

Information Retrieval (IR) is the field of study that examines how people find information and how tools (such as search engines and catalogues) can be constructed to help people find access information. Studies examine how the organization of information affects its retrieval, the types of searches people do, the kinds of search queries people can make effectively, and what determines the relevance of retrieved information. When information is available in enormous quantities and not clearly structured, people have difficulty finding relevant information and understanding important principles embedded in the information. The World Wide Web (WWW) is one example of Information Overload and its expansion has generated requirements for more effective access to global and corporate information repositories. These repositories are traditionally text based but increasingly include multimedia content such as audio (e.g. spoken language or music), graphics, imagery, and video. In text IR the user's requirements are expressed as text keywords and the query results is textual data in the form of word documents. In monolingual IR query and information to be looked for are encoded in the same language. The main question is how to retrieve relevant information from large text or hypertext collections automatically and intelligently.

2- Current approaches

Information retrieval, in its most simple form, is the process of gathering information on a particular subject. In its most basic terms, it is the process of matching a need to available knowledge. IR is a broad interdisciplinary and dynamic field that draws on many other disciplines. It stands at the junction of many established fields, and draws upon cognitive psychology, information architecture, information design, human information behaviour, linguistics, semiotics, information science, computer science and librarianship. Studies have typically approached IR from two major perspectives: from a rational approach which views IR as a mathematical model, as well as from a cognitive approach which views IR as an analysis of the process of information gathering done by people. In this sense, information retrieval systems not only include search engines, but also human constructed hierarchies, annotated bibliographies, and other specialized methods of presenting materials. Nevertheless, Search Engine technology and Automatic Text Information Retrieval have been fast-growing fields mainly due to the explosion of textual data available through the Web that renders inefficient the laborious task of human indexing. Therefore, statistical approaches have seen significant advances in recent years.

The main open issues in IR have to do with the Information access methods and the Information properties. Information access includes concepts such as information transmission and visualisation, categorisation and clustering, topic detection and tracking, summarisation, query formulation, information acquisition and extraction algorithms and their performance. The Information properties refer to the type of media data (text or multimedia), its structure (unstructured, semi-structured - XML, fully structured, hyperlinked – Web, mixture of types) and its heterogeneity (mono/multi-lingual, heterogeneous structures and services). Open issues regarding the heterogeneity of Information include the standardization

of non-trivial structures (e.g. Dublin Core) and services (e.g. XQuery text retrieval) and integration approaches based on uncertainty and vagueness.

Full text information retrieval is known to focus on the text itself, with semantics being handled in a rudimentary way. In traditional text retrieval the most straightforward way of locating the documents that contain a certain search term is to search all documents for the specified string (Full text scanning). Another well-known technique is the signature file approach. A fast text retrieval technique that is followed by many commercial systems is the inversion of the list keywords that represent the document content. A more sophisticated model than classical Boolean and Probabilistic models is the Vector Space Model (VSM) where a page is represented as a bag of keywords instead of a set of keywords as in the Boolean model. VSM takes frequency information into account. The language independent `bag-of-words' representations of documents have proved surprisingly effective for text classification. Common questions regarding term and document weighting schemes, normalisation, term stemming and common word elimination have been explored in depth in current bibliography. But the optimal representation of a text document remains an open research question. Some engines break documents and queries in phrases or even n-grams instead of words.

Recently, methods that try to capture more information about each document and achieve better performance have been researched and established in IR systems. These methods form three classes: (a) methods using parsing, syntactic information and Natural Language Processing (NLP) in general (b) algebraic methods based on dimensionality reduction techniques that extent the VSM, such as Latent Semantic Indexing (LSI) and (c) methods using neural networks and specifically spreading activation models.

Considerable advances have been made in recent years in syntactic modelling of natural language and development of efficient parsers with a broad domain. The task is to achieve automatic syntactic analysis and develop IR systems based on NLP. Progress is being made with syntax-directed semantic techniques such as lexical compositional semantics and with Artificial Intelligence techniques such as case frame analysis. But deeper semantic interpretation requires extensive knowledge engineering limiting the breadth of systems that depend on NLP.

Latent Semantic Indexing on the other hand has demonstrated improved performance over the traditional vector space techniques and has been successfully applied in many test IR systems. LSI, an optimal special case of multidimensional scaling, is a concept-based automatic indexing method that tries to overcome the two fundamental problems which plague traditional lexical-matching indexing schemes: synonymy and polysemy. It models the semantics of the domain in order to suggest additional relevant keywords and to reveal the “hidden” concepts of a given corpus while eliminating high order noise. The attractive point of this method is that it captures the higher order "latent" structure of word usage across the documents rather than just surface level word choice. This is done by modelling the association between terms and documents based on how terms co-occur across documents. Recently, Latent Semantic Analysis (LSA) has come under criticism, because its probabilistic model does not match the observed data. LSA assumes that words and documents form a joint Gaussian model. However, Gaussian models can generate negative values, and it is impossible to have a negative number of words in a document. Thus, a newer alternative is Probabilistic Latent Semantic Analysis (PLSA), based on a multinomial model, and is reported to give better results than standard LSA.

The data of today are electronically distributed and are represented in diverse formats and structures. Nowadays much emphasis is given in IR systems that have to deal with an excessive amount of unstructured or semi-structured data where no explicitly well-defined syntax for the documents in the archive exists. Because of the decentralized nature of its growth, the Web has been widely believed to lack of structure and organization as a whole. Even if web documents do share a syntax, there is no well-defined semantics associated with each syntactic component.

An open issue here is the size and coherence of the text repository from where we seek knowledge. At early years most of the research on information retrieval systems is on small well-controlled homogeneous collections such as collections of scientific papers or news stories on a related topic. Recently, the demand to find relevant information from large, noisy and non-homogenous corpora has become stronger. World Wide Web can be viewed as a graph, in which each node represents the page and edges connecting the nodes are the hyperlinks. The topology of this graph determines Web's connectivity and consequently how effectively can we locate information on it. The main goal of web IR is the automatic acquisition, indexing and ranking of documents in the Web. But, its enormous size, decentralized and dynamic nature and rapid growth pose a big challenge to search related pages for specific topic. Large-scale search engines struggle to cover the vast amounts of information that has been accumulated in the Web and maintain the freshness of their index. Furthermore, a very large portion of the web data is inaccessible through common web browsing or automatic crawling (hidden web). Recent studies (Kleinberg 2001) indicate that the Web contains a large, strongly connected core in which every page can reach every other by a path of hyperlinks. This core contains most of the prominent sites on the Web. The remaining pages can be characterised by their relation to the core. Due to good amount of resources for research in Web, many researchers are attracted into web IR area. There are many issues like extraction of the features from the pages, organizational structure of web, identifying community of pages, crawling the web, large-scale search engine and its architecture, web structure, personalized web search, page ranking methods, optimising web structure, web indexing etc. which are required for better web mining.

Information agents are programs that automatically perform customised information processing actions to deal with information overload problems. Examples of agents are the Web Crawlers which programs that traverse the hypertext structure of the Web automatically, starting from an initial hyper-document or a set of starting points (seeds) and recursively retrieving all documents referenced by that document. The recent trends in research of this field is the implementation of a focused crawler that intelligently avoids irrelevant portions of the web while visiting most relevant or promising pages early in the crawl process. This can help developing Vertical Search Engines that offer targeted and domain specific information to users. The open research problem is to efficiently reorder its crawl frontier (the queue of unvisited pages) when no content of the unvisited portion of the web graph is on hand.

What really differentiates hypertext from static text documents is the fact that the former, besides the text content, contain additional semantics, such as a document markup structure (Document Object Model – DOM), linking information that associates documents, citations and structured header (metadata) that precedes the relatively unstructured body. Link and social network analysis have been successfully applied both to academic citation data to identify influential papers and, more recently, to web hyperlink data to identify authoritative information sources. Recent techniques in web IR try to properly extract, exploit and integrate

all these features in order to efficiently process and acquire information so that distributed, portable, high-performance information processing engines can be developed. Clearly, outlinking information is available and can be used to implement well known relevance metrics and ranking algorithms such as HITS (Kleinberg 1998) and PageRank (Brin & Page), two of the most prominent algorithms in web IR. The heuristic underlying both of these approaches is that pages with many inlinks are more likely to be of high quality than pages with few inlinks, given that the author of a page will presumably include in it links to pages that s/he believes are of high quality. Lately, it has been shown that the ranking of the crawl frontier can be further improved by using the textual content from links that have been already visited. Numerous methods that try to combine textual and linking information for efficient URL ordering exist in the bibliography. Many of these are modifications, improvements or extensions of either PageRank (Mendelzon, Richardson & Domingos, Haveliwala) or HITS (Cohn & Hoffman). Chakrabarti et al. and Bharat & Henzinger also propose heuristic methods for differentially weighting links. Other algorithms such as SALSA (R. Lempel and S. Moran 2000), Spectral Filtering (S. Chakrabarti, B. Dom et al. 1998), HyCon (D. Mukhopadhyay et al. 2003) and Probabilistic HITS (Con & Chang 2000 - PHITS) are known to improve web search performance and provide quality pages.

Lack of domain knowledge means that user queries will inevitably have less satisfactory results. There are limitations to the amount of control an IR system has over their users' knowledge. Moreover success of query-oriented IR depends on the size of the query; short queries do not provide sufficient information to the IR system to distinguish relevant documents from irrelevant ones. But studies have shown that most of the queries consist of only a few keywords. On large scale libraries, especially over the Internet, user training is not an option to tackle with this problem. Thus, a system is needed that supports the user with additional domain knowledge. The approach taken is to refine, expand and re-weight the query automatically based on the documents retrieved by the original query. The common technique for automatic query expansion is to use pseudo-relevance feedback with top-K retrieved documents per query. There is need for distinguishing important terms and applying a proper weighting scheme. An alternative method is to expand each term in the original query with synonyms or related terms drawn from a generic on-line thesaurus. A third method to query expansion is based on interactive relevance feedback from the user. The system first returns a small number of matching documents; the user scans these, marking each document as "relevant" or "irrelevant". The system then uses this feedback from the user to formulate and launch a new query that better matches what the user is seeking.

Probably the most substantial evidence for automatic indexing has come out of the SMART Project (Salton 1966). The SMART system, developed at Cornell, is the one of the earliest IR systems that (1) use fully automatic term indexing, (2) perform automatic hierarchical clustering of documents and calculation of cluster centroids, (3) perform query/document similarity calculations and rank documents by degree of similarity to the query, (4) represent documents and queries as weighted term vectors in a term-based vector space, (5) support automatic procedures for query enhancement based on relevance feedback. SMART has been widely used as a testbed for research into, e.g., improved methods of weighting and relevance feedback, and as a baseline for comparison with other IR methods.

The Text Retrieval Conference (TREC) began in 1992 and serves as a major technology-transfer mechanism in the area of text retrieval. It attracts international participation from more than 100 research groups in retrieval technology, both from industry and academia. Its main goal is to accelerate the transfer of better text search and retrieval technology into commercial systems. Participating groups work with large, diverse test collections, submit their results for a common evaluation, and compare techniques and results.

3- Bibliography

J. Kleinberg, "Authoritative sources in a hyperlinked environment", In *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677, Jan. 1998.

S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", *WWW7 / Computer Networks*, 30(1- 7), pp.107-117, 1998.

K. Yang, "Combining text- and link-based methods for Web IR", In *Proc. Tenth Text Rerieval Conference (TREC-10)*. Washington 2002, DC: U.S. Government Printing Office.

S. Chakrabarti, M. Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery", *Computer Networks*, 31, pp. 1623-1640, 1999.

A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the Web", *ACM Transactions on Internet Technology*, 1(1), pp. 2-43, June 2001.

D. Cohn and T. Hoffman, "The Missing Link – A probabilistic Model of Document Content and Hypertext Connectivity", *Advances in Neural Information Processing Systems.* Vol.13, pp. 430-436, Boston, MA: MIT Press, 2001.

D. Bergmark, C. Lagoze, and A. Sbityakov, "Focused Crawls, Tunneling, and Digital Libraries". In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 91 – 106, 2002.

M. Richardson and P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank". *Advances in Neural Information Processing Systems.* Vol. 14 (pp. 1441-1448), 2002. Cambridge, MA: MIT Press.

R. Lempel and S. Moran. "The stochastic approach for link-structure analysis (SALSA) and the TKC effect". In *9th Int. WWW Conference*, Amsterdam, Nertherlands, May 2000.

S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan and Andrew Tomkins. "Topic Distillation and Spectral Filtering". *Artificial Intelligence Review*. Volume 13, Number 5-6, pp. 409-435, December 1999.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society for information Science*, 41, pp. 391-407, 1990.

E. Greengrass, “Information Retrieval: A Survey”. Online at http://www.csee.umbc.edu/cadip/readings/IR.report.120600.book.pdf

4- URLs

Text REtrieval Conference (TREC) http://trec.nist.gov
Special Interest Group on Information Retrieval (SIGIR) http://www.acm.org/sigir
CLEVER Project. IBM Almaden Research Center.
http://www.almaden.ibm.com/cs/k53/clever.html
Berkeley Digital Library SunSITE http://sunsite.berkeley.edu
Latent Semantic Indexing Web Site http://www.cs.utk.edu/~lsi
Reuters Corpus http://about.reuters.com/researchandstandards/corpus
CMU World Wide Knowledge Base (Web->KB) project
http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb
CMU Text Learning Group
http://www-2.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www
Apache Jakarta Lucene search engine
http://jakarta.apache.org/lucene/docs/index.html

Cross Language Information Retrieval

Rauber/TU Wien-IFS

1- Description

Cross-language information retrieval (CLIR) describes the task of finding documents written in one language with queries formulated in a different language. Basically, there are two different approaches, namely translation of the documents and translation of the queries. A third approach is the translation of both the query as well as the documents into an intermediary representation, e.g. latent semantic indexing (LSI). For a good overview of the state of the art of the various techniques employed in handling multi-lingual document collections, see (Hovy *et al*, 2001; Grefenstette, 1998).

2- Current approaches

Most approaches currently follow the query-translation principle due to memory efficiency reasons, as the source documents are stored only once. Query terms are translated using machine translation, to retrieve documents in a language other than the query language. Yet, this usually has the disadvantage of word sense ambiguity caused by the lack of context a translation engine could use (Hull & Grefenstette, 1996). Especially short queries, which are most commonly issued by users of search engines, suffer from bad automatic translation. Dictionary-based methods combined with query expansion techniques (Ballesteros & Croft, 1997) or structured translation (Sperer & Oard, 2000) try to reduce the ambiguity of the translated query and therefore, to increase retrieval performance. Another technique is using parallel text corpora to select the most appropriate query terms from the set of possible translations.

A different, yet essential, approach to access a multilingual document collection is interactive exploration. However, only little research work has been reported in this field so far. A different approach to organizing multilingual document collections using Self–Organizing Maps (SOMs) is described in (Lee & Yang, 2000), reporting on some initial experiments comparing SOM clustering performance on a small parallel Chinese – English corpus. This work is continued in (Lee & Yang, 2003) where document collections of between 40 and 100 documents in English and Chinese are analyzed by clustering words rather than documents, obtaining concept clusters. More recently, work employing sentence clustering before the translation process in a document summarization system has been reported in (Chen et al. 2003). Documents are partitioned into event clusters, from which summaries are subsequently created. For an extensive list of literature on this subject, see the proceedings of the annual cross-language information retrieval evaluation forum workshop (CLEF, 2002) and the Japanese sponsored NTCIR competitions.

3- Bibliography

Ballesteros, Lisa and B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR97)*, pages 84 – 91, Philadelphia, PA, 1997.

Chen, H.-H, and J.-J. Kuo, and T.-C. Su. Clustering and visualization in a multi-lingual multi-document summarization system. In *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR 2003)*, number 2633 in LNCS, pages 266–280, Pisa, Italy, April 14-16 2003. Springer.

CLEF: Cross language evaluation forum. http://clef.iei.pi.cnr.it:2002

Grefenstette, Gregory, editor. *Cross-Language Information Retrieval*. Boston: Kluwer Academic Press, 1998.

Hovy, E. and N. Ide, R. Frederking, J. Mariani, and A. Zampolli, editors. *Multilingual Information Management: Current Levels and Future Abilities*, volume 14-15 of *Liguistica Computazionale*. Insituti Editoriali e Poligrafici Internazionali, Pisa, Italy, 2001. http://www-2.cs.cmu.edu/~ref/mlim/.

Hull, David A. and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR96)*, pages 49 – 57, Zürich, Switzerland, 1996.

Lee, C.-H. and H.-C. Yang. Towards multilingual information discovery through a SOM based text mining approach. In T. Ah-Hwee and P. Yu, editors, *Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000)*, pages 80–87, Melbourne, Australia, August 28 - September 1 2000. Deakin University, Australia.

Lee, C.-H. and H.-C. Yang. A multilingual text mining approach based on self-organizing maps. *Applied Intelligence*, 18(3): 295–310, May/June 2003.

Sperer, R. and D.W. Oard. Structured translation for cross-language information retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR00)*, Athens, Greece, 2000.

4- URLs

CLEF: cross language evaluation forum. Website. http://clef.iei.pi.cnr.it:2002

NTCIR: http://research.nii.ac.jp/ntcir/workshop/

Douglas Oard's CLIR and MT resource page: http://www.glue.umd.edu/~oard/research.html

Text Classification

Novovicova/UTIA

1- Description

Text document classification task can be described as follows: Given a finite set of predefined document classes and a finite set of documents (training documents), construct a classifier which, given a new (free) document, finds the class or classes to which the new document ought to be assigned. The classes are just symbolic labels, and no additional knowledge of their meaning is usually available. It is often the case that no data such as e.g. publication date, document type, publication source is available either. In these cases, classification must be accomplished only on the basis of knowledge extracted from the documents themselves. Depending on the application, text classification may be either a single-label task (i.e., exactly one class from the set of predefined classes must be assigned to the document) or a multi-label task (i.e., any number of classes may be assigned to the document). Text classification is highly language dependent. Nearly all published results are related to English only. The morphological analysis for some languages (e.g., Slavic, German) is much more difficult than for the English language and the corresponding research is in its early stages yet

2- Current approaches

The automated classification of text documents into predefined classes has gained a prominent status in the information system fields in the last ten years, due to the increased availability of documents in digital form and the need to organize them. Document classification may appear in many applications including e-mail filtering, mail routing, spam filtering, news monitoring, selective dissemination of information to information consumers, automated indexing of scientific articles, automated population of hierarchical catalogues of Web resources, identification of document genre, authorship attribution, survey coding, and so on. Automated text classification is attractive because manually organizing text document bases can be too expensive, or simply infeasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems overcomes that of trained human professionals, thanks to a combination of information retrieval techniques and machine learning techniques.

In the research community the dominant approach to text classification problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning from a set of preclassified documents. We can roughly distinguish three different phases in the design of a text classification system: document indexing, classifier construction, and classifier evaluation. Some of the actual techniques for dealing with the tasks of document indexing and classifier construction will be now described. The overview of Sebastiani (Sebastiani 2002) discusses the main approaches to text classification.

Document indexing denotes the mapping a document into a compact representation of its content. Many document-indexing methods usually employed in text classification are from information retrieval area. The indexing proceeds in four steps: morphological analysis, elimination of non-significant words, frequency analysis and index term weighting. A text

document is typically represented as a vector of term (word) weights (also known as features) from a set of terms (called dictionary) that occur at least once in at least *k* (chosen number) documents. A common characteristic of text data is its extremely high dimensionality. The number of potential features (several tens of thousands) often exceeds the number of training documents. Dimensionality reduction is a very important step in the text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy.
Dimensionality reduction often takes the form of feature selection. Methods for feature subset selection for text classification task use some evaluation function that is applied to a single feature. All features are independently evaluated, a score is assigned to each of them and the features are sorted according to the assigned score. Only the highest scoring terms are used for document representation. Scoring of individual features can be performed using some of the measures, for instance, document frequency, term frequency, mutual information, information gain, odds ratio, chi squared-statistics, term strength. In the paper of Forman (Forman 2003) is presented an extensive comparative study of twelve feature selection criteria for the high-dimensional domain of text classification. Recent study (Novovicová et al. 2004) proposed to use sequential forward selection method based on improved mutual information as a criterion for reducing the dimensionality of text data. This feature evaluation function takes into consideration how features work together.
Alternatively, dimensionality reduction may take the form of feature extraction: a set of new terms is generated from the original term set in such a way that the newly generated terms are more stochastically independent from each other than the original ones were. The feature extraction methods such as latent semantic indexing or term clustering (Slonin et al. 2001) are used in text classification. Recent work on feature extraction methods is focused on methods specific to problems in which training data exist, i.e. on supervised feature clustering techniques, which have shown better performance than unsupervised techniques. Feature/word clustering is a powerful alternative to feature selection for reducing the dimensionality of text data. Dhillon et al. have presented (Dhillon et al. 2003) a new information-theoretic divisive algorithm for feature/word clustering and applied it to text classification. This algorithm minimizes the within cluster Jensen-Shannon divergence and simultaneously maximizes the between-cluster Jensen-Shannon divergence. It is shown that the divisive clustering algorithm is an effective technique for building smaller class models in hierarchical classification.

The number of categories of classifier learning techniques that have been used in text classification includes the probabilistic approaches, decision tree and decision rule classifiers, regression methods, neural networks, batch and incremental learners of linear classifiers, example-based methods, support vector machines, genetic algorithms, hidden Markov models, and classifier committees (which include boosting methods), maximum entropy modelling. The support vector machines and boosting are two methods that have shown the best performance in comparative text classification experiments performed so far. The support vector machines method has been introduced in text classification by Joachims (Joachims-1998) and subsequently used in several other text classification works.

In text classification research, the experimental evaluation of classifier usually measures its effectiveness rather than its efficiency, that is its ability to take the right classification decisions. Classification effectiveness is usually measured in terms of the classic information retrieval notions of precision and recall. Measures alternative to precision and recall commonly used in the machine learning literature, such as accuracy and error are not widely used in text classification.

Recently, text classification research is pointing in several directions. One of them is the attempt at finding better representations for text; while the bag-of-words model is still the unsurpassed text representation model, researchers have not renounced to the belief that a text must be something more that a mere collection of tokens, and that the quest for models more sophisticated than the bag-of-words model is still worth pursuing.
A further direction is investigating the scalability properties of text classification systems. Real applications of text classification often require a system with tens of thousands of classes defined over a large taxonomy. Although many classification methods have been published, it is difficult to tell which one would scale to applications with such a large number of classes (Yang et al. 2003) . A part of the difficulty comes from the fact that many of them were evaluated using a small number of classes.
The attempts are also at solving the labelling bottleneck, i.e. at coming to terms with the fact that labelling examples for training a text classifier when labelled examples do not previously exist, is expensive. As a result, there is increasing attention in text classification by semi-supervised machine learning methods, i.e. by methods bootstrap off a small set of labelled examples and leverage on unlabelled examples too (Nigam et al. 2000).

Finite mixture models have been employed in a number of text processing applications, such as text classification (e.g. Juan et al. 2002, Ueda 2003, Novovicová et al. 2003) indicating that they are powerful for text processing. The usage of finite mixtures for class-conditional probability functions is a useful method, because mixture models are able to represent arbitrarily complex probability functions. The mixture approach to learning on text document is based on the fact that documents in the same class are often mixtures of multiple topics. Mixtures are flexible enough for finding appropriate tradeoffs between model complexity and the amount of the training text data available. Usually, model complexity is controlled by varying the number of mixture components while keeping the same parametric form for all components.

Automated text classification has evolved into a fully blossomed research field, which has delivered workable solutions that have been used in a wide variety of real-world application domains.

3- Bibliography

G. Forman. An experimental study of feature selection metrics for text categorization. Journal of Machine Learning Research, 3: 1289-1305, 2003.

I.S. Dhillon, S.Mallelam, and R.Kumar. A divisive information-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research, 3: 1265-1287, 2003.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the ECML'98, 137-142, 1998.

A. Juan and E. Vidal. On the use of Bernoulli mixture models for text classification. Pattern Recognition, 35: 2705-2710, 2002.

K. Nigam, A.K. McCallum, S.Thrun, T. Mitchell. Text classification from labelled and unlabelled documents using EM. Machine Learning, 39: 103-134, 2000

J. Novovicová and A. Malík. Application of multinomial mixture model to text classification. Pattern Recognition and Image Analysis, Lecture Notes in Computer Sciences 2652, Springer-Verlag, Berlin, 646-653, 2003.

J. Novovicová, A. Malík, and P. Pudil. Feature selection using improved mutual information for text classification. Structural, Syntactic and Statistical Pattern Recognition, Lecture Notes in Computer Science 3139, Springer-Verlag, Berlin, 2004.

F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

L.D Baker. and McCallum, A.K Proc. of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, eds. W.B. Croft, A. Moffat, C.J.V. Rijsbergen, R. Wilkinson & J. Zobel, ACM Press, New York, US: Melbourne, AU, 96–103, 1998.

N. Slonim and N. Tishby. The power of word clusters for text classification. Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research, Darmstadt, DE, 2001.

N. Ueda and K.Saito. Parametric mixture models for multi-labeled text. Proc. of Neural Information Processing Systems, 2002.

Y. Yang, J. Zhang, and B. Kisiel. A scalability analysis of classifier in text categorization. Proc. of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, 96-103, 2003.

4- URLs

CMU Text Learning Group http://www-2.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www
CMU World Wide Knowledge Base (Web->KB) project
http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb
Special Interest Group on Information Retrieval (SIGIR) http://www.acm.org/sigir
Reuters Corpus http://about.reuters.com/researchandstandards/corpus
Newsgroups data http://www.cs.cmu.edu/