MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 11
December 2000

The *Machine Translation Review* incorporates the Newsletter of the Natural Language Translation Specialist Group of the British Computer Society and appears twice yearly.

The Review welcomes contributions, articles, book reviews, advertisements, and all items of information relating to the processing and translation of natural language. Contributions and correspondence should be addressed to:

Derek Lewis
The Editor
Machine Translation Review
School of Modern Languages
Queen's Building
University of Exeter
Exeter
EX4 4QH
United Kingdom

Tel/fax: +44 (0)1392 264296
E-mail: D.R.Lewis@exeter.ac.uk

ISSN 1358–8346

Please note: From April 2000 this Review will be published electronically and will be available on our web site at the British Computer Society (see page 5). The format will be in HTML, in the same way as some back copies have already been stored electronically, so it will be easy for readers to print copies if they wish. Each section will be separate so readers may print selected parts only.

Some copies will be printed for the Copyright libraries and for purchase at a modest price plus postage and packing for those without electronic access. Members of the Natural Language Translation Specialist Group of the British Computer Society will be advised of each issue.

Contents

Group News and Information

Letter from the Chairman

The British Computer Society
Charity No. 292786

THE NATURAL LANGUAGE TRANSLATION SPECIALIST GROUP

72 Brattle Wood
Sevenoaks
Kent, TN13 1QU
Tel: 01732 455446
Office: 0207 815 7472
Fax: 0207 815 7550
E-mail: wiggjd@sbu.ac.uk

18 January 2001

'I regret that we have been unable to produce another MT Review until now as it has not been possible with our limited resources to publish another issue as well as organise the MT Conference in the Autumn.

I'm happy to say that our fourth International Machine Translation Conference, MT 2000, held at the Crossmead Conference Centre at Exeter University under the calm and efficient guidance of Derek Lewis at the University, was a great success and the 60 various attendees were well served both in educational and culinary terms. Copies of the Proceedings are now available from Derek Lewis at £20 for members (please see our website at http://www.bcs.org.uk/siggroup/sg37.htm for details of cost of postage etc.).

We also have to thank Professor Ruslan Mitkov from Wolverhampton University for his invaluable contribution in organising the paper reviewing processes and for putting the programme together.

There are some changes afoot in the Group. As you will already know, this Review is being published first on our website from which printed copies can be obtained easily. However, we shall produce some printed copies for the Copyright Libraries and for sale at a modest price of £2.00 for those who would like printed copies (please see our website for details of cost of postage etc.).

What you will not know is that we have regretfully decided to reduce the number of issues from two to only one per year.

We have also decided to transfer our mailing list which is maintained at the British Computer Society Headquarters to an automated e-mail service elsewhere. We will be contacting our membership about this in the near future.

Once again we must ask you to consider contributing to this Review. We would still welcome more articles, papers and reports on the subject of machine translation and related subjects such as computer assisted language teaching, computer based dictionaries and aspects of multilinguality in computing etc. We would welcome papers from staff and

students in linguistics and related disciplines, and from translators and any other users of MT software.

May I remind members yet again, that they do not need to live near London to assist the Committee. We do not have sufficient funds to pay travel expenses for all Committee members to attend meetings, but we still welcome Correspondent members. Correspondent committee members are otherwise treated as full members of the committee and kept advised of all committee business. Anyone interested in helping should contact me or any other Committee member.

Our committee still requires a treasurer, although in our case the role is more of an auditor since all our transactions are processed by the BCS. This post does, of course, require some knowledge of accounting, but not much I'm glad to say, and, as mentioned above this does not need to be for someone in the London area. Anybody interested to know more, please contact me.

All opinions expressed in this Review are those of the respective writers and are not necessarily shared by the BCS or the Group.'

The Committee

The telephone numbers and e-mail addresses of the Committee are as follows:

David Wigg (Chair)	Tel.: +44 (0)1732 455446 (H)
	Tel.: +44 (0)207 815 7472 (W)
	E-mail: wiggjd@sbu.ac.uk
Monique L'Huillier (Secretary)	Tel.: +44 (0)1276 20488 (H)
	Tel.: +44 (0)1784 443243 (W)
	E-mail: m.l'huillier@rhbnc.ac.uk
Derek Lewis (Editor)	Tel.: +44 (0)1404 814186 (H)
	Tel.: +44 (0)1392 264296 (W)
	E-mail: d.r.lewis@exeter.ac.uk
Douglas Clarke	Tel.: +44 (0)1908 373141
Ian Kelly	Tel.: +44 (0)1276 857599
	E-mail: idkk@iddk.com
Veronica Lawson	Tel.: +44 (0)207 7359060
	E-mail: veronica_1@compuserve.com
Roger Harris (Webmaster)	Tel.: +44 (0)208 800 2903 (H)
	E-mail: rh@nationalfinder.com
Correspondent Members:	
Gareth Evans (Minority Languages)	Tel.: +44 (0)1792 481144
	E-mail: g.evans@sihe.ac.uk
Ruslan Mitkov	Tel: +44 (0)1902 322471 (W)
	E-mail: R.Mitkov@wlv.ac.uk

BCS Library

Books kindly donated by members are passed to the BCS library at the IEE, Savoy Place, London, WC2R 0BL, UK (tel: +44 (0)207 240 1871; fax: +44 (0)207 497 3557). Members of the BCS may borrow books from this library either in person or by post. All they have to provide is their membership number. The library is open Monday to Friday, 9.00 am - 5.00 pm.

Website

The website address of the BCS-NLTSG is: http://www.bcs.org.uk/siggroup/sg37.htm

First Steps of Language Engineering in the USSR: The 50s through 70s

by

Michael Blekhman and Boris Pevzner

Michael S. Blekhman: 4850 Edouard Montpetit, #19, Montreal, QC, Canada.
E-mail: ling98@canada.com. Internet: www.ling98.com
Boris Pevzner: Levy Eshkol 7/51, Azorim Netanya, Israel 42463.
E-mail: ling98@canada.com. Internet: www.ling98.com

Keywords: language engineering, machine translation, information retrieval, cybernetics, USSR.

Abstract

Language engineering was one of the most interesting fields of science and technology in the former Soviet Union. However, few people (if any) know details of its development. The authors of the present paper have been and are active in machine translation, computational lexicography, information retrieval, automatic abstracting and indexing. Boris Pevzner began practical work in language engineering in the 1960s, while Michael Blekhman, Dr. Pevzner's pupil, became a researcher in mid 1970s. Thus we can provide analysis 'from inside'.

In 1954, a year after Joseph Stalin's death, the Soviet Union 'rehabilitated' cybernetics, which had been considered an 'imperialist' science before that. The beginning of the new era in science was marked with Acad. Solodovnikov's public lecture on cybernetics given in the Polytechnic Museum in Moscow. Several more years passed since then before those fundamental ideas laid in the foundation of cybernetics were understood and appreciated by the young generation of scientists.

One of the results of this understanding was the appearance of Russian translations of classic books on cybernetics and applied mathematics. The translations were made by the leading Soviet researchers V. Cherniavsky, D. Lakhuti, A. Yesenin-Volpin, and some others. Papers on theoretical issues of machine translation were published regularly in the issues of *Masshinnyi perevod* ('Machine Translation'). They presented translations of papers by the most prominent linguists and cyberniticians, Noam Chomsky among them.

Centers of research and development in this field were the All-Union Institute for Scientific Information - VINITI (headed by Prof. A. Mikhailov) and Laboratory for Electric Modeling (headed by Prof. A. Vasilyev).

Original Russian publications appeared at the same time, the most widely readable edition being *Problemy kibernetiki* ('Problems of Cybernetics'). It was there that various aspects of automatic text processing were discussed for the first time in the Soviet Union. In particular, papers by O. Kulagina and N. Moloshnaya covered some practical issues of machine translation.

In the 50s and 60s, I. Belskaya suggested a detailed algorithm of automatic English-Russian translation. The algorithm was published as a large two-volume book.

All those break-through ideas were not tested on representative text corpora, however. An illusion existed that they were absolutely logical and non-contradictory and would lead to quick and efficient problem solution. As one of the linguists pointed out, having large

dictionaries and detailed traditional grammars, one only needed a powerful computer to develop real-life text processing systems.

In the late 50s, L. Gutenmacher published one of the first Russian monographs on cybernetics – *Information Retrieval Systems*. He discussed various aspects of information retrieval, such as software, hardware, and communication through telephone channels. Gutenmacher foresaw the time when one would be able to access remote libraries from home via telecommunication channels (a prototype of the internet).

G. Lesskis, a Moscow-based linguist and literature researcher, carried out a deep statistical analysis of a large fiction text corpora and found out an extremely interesting phenomenon: the lengths of sentences vary depending on the narration phase: introduction, culmination, and epilogue.

In the early 60s, two large scientific conferences on automatic text processing were held in Moscow: 'Problems of Semiotics' at the Institute of Foreign Languages, and 'Problems of Automatic Information Retrieval' at VINITI. Many researchers took part in both. One of the sessions at the former was presided by Acad. A. Markov, the prominent scientist who created an algorithmic model, descriptive enough for linguistic analysis. The algorithm was formulated in terms of word conversion (the Markov algorithm). A broad spectrum of problems was approached dealing with algorithmic text processing:

- semiotic analysis of fiction, formalization of the ties between the characters;
- formalisms describing sense relations in texts;
- morphological, syntactic, and semantic analysis;
- formal languages, both well-known and 'exotic' ones, such as the language of the cinema (S. Genkin), and many other problems.

The conference on automatic retrieval approached both theoretical and purely practical issues. Quite a number of contributions were made by young researchers who worked in the laboratories headed by G. Vlaeduts and V. Cherniavsky.

One of the most important events was developing the world's first personal computers of the *Mir* series at the Institute of Cybernetics of the Ukrainian Academy of Sciences in Kiev. Some time later Acad. V. Glushkov, director of that institute, published his breakthrough monograph named *Paperless Informatics,* suggesting electronic data processing and exchange (a prototype of modern information technologies).

In the mid-60s, two exhibitions on information technology were held in Moscow. The contributions, made by the members of the Council for Economic Cooperation (socialist countries), embraced operational models rather than commercial products: second generation computers *Minsk* and *Ural*, photocopying machines, office stationery, information retrieval systems based on those machines. Some stands displayed Western technology, although no Western representatives attended the exhibitions.

One of the most remarkable events was Norbert Wiener's visit to Moscow and his lectures at Moscow State University. The great scientist was met with understanding and enthusiasm from the young Russian students and experienced researchers. He appreciated the high professional level of the audience.

Also in the 60s, VINITI began publishing what became the most authoritative and respectable Soviet journals in the field of information processing and language engineering, the monthly *Nauchno-technicheskaya informatsiya* ('Scientific and Technical Information'). Since the first issue, its readers and contributors have been both practical researchers and theoreticians from all over the USSR.

At the same time, a theoretical department, 'Semiotics,' was created at VINITI. Its main fields of research were formal grammars, multiple-valued logics, information retrieval, and formalized methods of text processing. A scientific council and a postgraduate center for scientific and technical information were established at VINITI at the same time.

In the same period of time, a fundamental monograph was published by A. Mikhailov, A. Chernyi, and A. Gilyarevskyi: *Foundations of Informatics*. It became a textbook for thousands of students in the Soviet Union.

In the mid-60s, information technologies gained a great momentum in the USSR. All-Union, Republican, departmental, and local information centers were set up. Their main task was creating, processing, and distributing textual and factual databases as well data-processing technologies. The databases comprised millions of entries. Search mechanisms were based on keywords and topic categories. The All-Union system of scientific and technical information was unprecedented in the world. It had a layered multidivisional structure and served thousands of institutions all over the USSR. Data exchange was carried out using magnetic tapes. Telecommunications were also introduced.

Machine Translation in the USSR

This topic deserves special discussion. Due to the lack of space, however, we will only shed light on a few points of thousands.

The Moscow State University group headed by O. Kulagina and N. Moloshnaya developed algorithms for automatic morphological and syntactic analysis.

At VINITI, a group led by Yu. Shreider developed algorithms for recognizing proper names in real-life texts. The algorithms were based on calculating distances between words in the texts. The group focused attention on theoretical aspects of machine translation.

Researchers led by I. Mel'chuk at the Institute of Foreign Languages pioneered in using a semantic component in machine translation. That component was introduced as a explanatory-combinatorial dictionary. Each word-entry in the dictionary was made according to a universal scheme using the so-called lexical and logical functions. Such an approach was supposed to provide complete description of the word-entry including phrases and idioms.

Research and developments in the machine translation area were also carried out at the All-Union Patent Institute, Leningrad State University, and some other centers.

The most outstanding person to mention is, to our opinion, Raimund Piotrowski, professor at the Leningrad Pedagogical Institute, a man whose role in Soviet language engineering has been really great. He is both a brilliant linguist and a very energetic organizer. In the early 1970s, he founded the All-Union linguistic group, which he called Statistica Rechi ('Speech Statistics'). It united language engineers from all over the USSR: Leningrad, Moscow, Ukraine, Kazakhstan, Moldavia, Uzbekistan, Azerbaijan, etc.

The first operational Soviet MT system was developed in 1976 at Chimkent Teachers Training College in Kazakhstan, by the Kazakhstan subgroup, Speech Statistics, headed by Prof. K. Bektayev and Prof. P. Sadchikova. The system ran on IBM-compatible mainframes and performed word-for-word and phrase-for-phrase English-Russian translation of patent chemical texts. The system was used at the Institute of Chemistry, Kazakhstan Academy of Sciences.

Piotrowski's Moscow colleague, Prof. Yuri Marchuk, Director of the All-Union Center for Translations, headed a project covering three MT systems: English-Russian (AMPAR), German-Russian (NERPA), and French-Russian (FRAP). The AMPAR system was launched in 1977. It was used for generating raw translations of technical texts both at the Center and at some departmental research institutes. Marchuk published a two-volume English-Russian contextual dictionary that can be used (and I plan to use it!) for disambiguation purposes. Dr. Yevgeni Lovtski developed a special language for representing linguistic rules in AMPAR. Doctors Boris Tikhomirov, Zoya Shaliapina, and Nina Leontieva investigated various aspects of semantic-based and transfer-based MT. I believe that Zoya was the leading expert in Japanese-based MT in the USSR.

Dr. Boris Pevzner published in the early 70s a series of papers on example-based text processing.

The 70s were a period of scientific confrontation of two conceptions: the practical ('engineering') approach to machine translation, most vividly expressed by Raimund Piotrowski, and the theoretical approach, backed by such outstanding linguists as Prof. Igor Melchuk and Prof. Yuri Apresian. They opposed the idea of **automatic** translation (AT) to Piotrowski's **machine** translation (MT), and argued that the linguist's task is to offer an in-depth description of the language as the foundation of an AT algorithm instead of gradually improving an imperfect MT system. Apresian's group developed the ETAP family of pilot MT systems translating from French and English into Russian. It is interesting that the word-for-word English-Russian translation module was used for translating patent titles in the INPADOC patent information retrieval system.

However, it was in the 1990s, with the advent of personal computers, that machine translation was made accessible to hundreds of thousands of end users. Would it have been possible without the first steps made by the pioneers?

Keeping Translation Technology under Control

by

Dawn Murphy

Introduction

The following is a summary of a talk which I gave to the BCS Natural Language Translation Group on 8 December 1999. The talk was on the subject of controlled authoring and how better results from machine translation and translation memory tools can be achieved by restricting the language of the source text.

A new era for translation technology

Translation technology is becoming more widely accepted these days in the global marketplace. All companies who produce technical documentation are under increased pressure to reduce translation costs as they enter new markets, and to reduce time-to-market as global markets require virtually simultaneous release. While in the past they may have been sceptical about using translation technology, now many companies are sceptical of translation providers who do not use such tools.

Many companies nowadays expect their translation providers to use translation memory tools to cut translation costs. However, often these same clients are disappointed when the level of text re-use (100% matches) reported by the translation provider is much lower than they had expected. On analysis, it is often found that the lower hit rate is down to the way the source text is written. In effect, the source text is rewritten every time, so not surprisingly, it has to be retranslated every time.

As well as translation memory tools, fully automated machine translation is starting to come of age, with successful implementations being reported at various sites. The success of these implementations is due to various factors, including the suitability of the subject domain and text types and the customisability of the system. The quality of the source text is one of the most vital criteria for successful use of MT at present. The more restricted the source language, the less room for inaccuracy in translation. This is where controlled authoring comes in. Controlled authoring is not a new concept, having been discussed, and used, for a number of years, both as a means of improving English texts and as a means of gaining more control over the output of MT. It's gone through some rough times too, with many people questioning whether it is worth the investment.

The internet raises the stakes

What is happening now though is that we are reaching a point in the translation industry, where translation technology is becoming an essential part of the translation process rather than a curious experiment. The emergence of internet-based translation services aimed at corporate users is raising expectations too, with the promise of fast turnarounds for translation jobs. People are starting to expect everything at ‘internet speed’, i.e. translation at the touch of a button.

In addition, companies with a global presence are starting to realise the importance of localising their website. This is creating a huge amount of translation work which needs to be turned around virtually immediately, as website content may be updated on a daily basis. The

volumes and turnaround times involved are often so high that traditional translation methods just cannot keep up with the demand, or the cost becomes unpalatable.

The application of MT and TM in both these contexts has obvious potential, ensuring rapid turnarounds and lower costs, though it should be said that MT is not always suitable for the more marketing-oriented content. Controlling the source text is suddenly more important than ever if we are to make the most of the translation technologies available to us and if quality translations are to be delivered on time. The investment is now balanced against a much higher return.

Controlling the source

Controlled authoring involves both restricting the language used by authors and using appropriate tools to ensure maximum re-use of existing text. Maximum re-use could be achieved through the use of author memory technology, which would prompt the author to phrase a single idea in the same words every time. With multilingual content management databases, this re-used text can be directly linked to its translation and thereby bypass both MT and TM for 100% matched text, bringing in TM for similar text (fuzzy matches) and MT for new text where appropriate.

Restricting the language which is used by authors essentially boils down to avoiding ambiguous constructions and ensuring that the correct terminology is used. This process can then be supported by look-up and checking tools which the author can use to ensure they do not use any disallowed constructions. Typical rules might be: 'Avoid using pronouns', or 'Keep to one instruction per sentence'. Criticisms often levelled at controlled language are that it produces unnatural, stilted text, but this is not necessarily the case. Most translators will be able to point you to a text they have been given to translate which was virtually incomprehensible in the source language. Carefully trained authors and carefully designed controlled languages can often improve the source text, making it easier to read and easier to follow instructions. In fact, the first controlled languages were designed for this very purpose, and it was only afterwards that the possibilities for translation were considered.

Terminology

Consider the following examples of how a term can be written in different ways, all taken from the same document:

front right hand wheel arch
front right-hand wheel arch
front rh wheel arch
front right wheel arch
right hand front wheel arch
right-hand front wheel arch
rh front wheel arch

Apart from confusing a translator who might think these terms referred to different concepts, this uncontrolled use of terminology can lead to sentences not being matched in translation memory tools. MT usually relies on user-customisable dictionaries to translate terms correctly so the only way to make sure that all variants of a term are translated are to add each one to the dictionary. But who can say that the next author will not write 'right front wheel arch', yet another variant? And how many separate entries will be needed to code this many variants for each term?

Controlled language

In the following example, the same idea is phrased in four different ways. This would ideally be avoided by using an author memory tool but if not, then controlled language can be applied to standardise the way an idea is phrased.

Center the steering wheel and lock in position.
Center the steering wheel. Lock in position.
Center the steering wheel. Lock it in position.
Center the steering wheel. Lock the steering wheel in position.

In the second line the sentence has been split into two so that there is one instruction per sentence. This is particularly important for translation memory tools, as re-use is much more likely with short simple sentences expressing one idea than with longer more complex sentences. In a software manual, the instruction 'Click on OK' might appear several times. It is much less likely that 'Enter your license key number and click on OK' will appear more than once in the same manual. In addition, the 'and' which links the two instructions is ambiguous in itself. It could mean 'and then' implying consecutive actions, or it could mean 'while' implying simultaneous actions. In the third line, the object has been made explicit so there is no ellipsis in the sentence. Where the object is not made explicit, the instruction could be incorrectly interpreted by a MT engine, which might see 'lock in' as a phrasal verb or even a noun followed by a preposition. In the fourth line, the object has been changed from a pronoun to the complete noun, eliminating referential ambiguity, and thereby reducing the potential for incorrect translation by an MT system.

Obtaining good results from translation memory relies on a restricted use of language. If the same idea is expressed the same way every time because the controlled language rules guide the author to expressing an idea in a certain way, the hit rate from TM should be higher.

Summary

Using controlled language creates unambiguous text, which presents MT with far fewer opportunities for mis-translation. Standardised terminology means fewer unrecognised words for an MT system, and more hits from TM. More consistent text makes MT easier to tune, by building up the dictionaries and tuning the grammar rules to suit the relatively small set of constructions used by the authors, and again, means more hits from TM because there are fewer ways to say the same thing. Finally, even before you reach the controlled language stage, using content management and author memory tools to maximise re-use of existing text will help to solve many of the cost, quality and time-to-market problems faced by companies today in the production of multilingual documentation.

(At the time of Dawn's talk, she was a language engineering consultant with Multilingual Technology Ltd (MTL). MTL has since been acquired by Berlitz GlobalNET to become the Berlitz GlobalNET Solutions Group, offering consulting and solutions for web globalisation. See the press release at http://www.berlitzglobalnet.com/english/press/the_press_room.asp. Dawn continues to work as a technical consultant with Berlitz GlobalNET, focusing on translation technology and multilingual content management.)

Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems

by

Natalia V. Loukachevitch , Boris V. Dobrov

louk@mail.cir.ru; dobroff@mail.cir.ru

Centre for Information Research, 339, Scientific Research Computer Centre of

Moscow State University Vorobyevy Gory, Moscow, Russia

Abstract

The paper describes the technique of construction of a structural thematic summary. A structural thematic summary represents contents of texts by indication of the main theme and subthemes of a text simulated by sets of terms corresponding to these themes. A structural thematic summary comprises the most informative fragments of thematic representation of a text that contains all terms of the text divided to thematic nodes**.** A structural thematic summary is created on the basis of detailed bilingual description of the sociopolitical domain. It can represent contents of documents of any size and different genres. Language of documents and corresponding structural thematic summaries can be Russian or English.

1. Introduction

In multilingual information retrieval there is a serious problem of how users can estimate the relevance of retrieved documents and how they can choose the most relevant documents for computer or human translation.

Summarisation of texts is usually considered as one of important tools helping to evaluate relevance of texts to users' information needs. Summarisation of texts in broad domains or domain-independent summarisation is mostly based on passage extraction (Salton 1989). Such summaries are constructed as ordered sequences of sentences or paragraphs of initial texts chosen using some criteria**.** Users of multilingual systems can be unfamiliar with the language of document collection. Machine translation of retrieved documents or their summaries can significantly slow down the process of the choice of relevant documents.

Translations of the most frequent words can be used as representative lists of contents of documents. But in these lists the terms corresponding to different topics of texts can be intermixed. Also manifold terms of the most frequent topic can occupy all the available space, and terms of other important topics will be missed.

Boguraev el al (1997) proposed the dynamic presentation of document content based on salient fragments of sentences, but to receive qualitative translations of such fragments into other languages is as difficult as it is to translate the whole text.

In this paper we describe the construction of the structural thematic summary of a text. The structural thematic summary describes contents of texts by a representation of the main theme of a text simulated by sets of terms corresponding to the main theme. Construction of the structural thematic summary is based on the domain-specific thesaurus, specially created as a

linguistic resource for automatic text processing. The structural thematic summary is the most informative fragment of the thematic representation of a text that is a result of complicated process of thematic analysis of texts including term disambiguation and analyses of cohesion relations.

Construction of thematic representation is based on the Russian-English Thesaurus on Sociopolitical Life specially created as a tool for automatic text processing. The Thesaurus contains 24 thousand concepts, 50 thousand terms and 90 thousand relations between concepts.

Thematic representation including the terms of a text is too detailed to serve as a summary for users of information systems. So we construct a structural thematic summary containing terms of the main theme and important subthemes of a text as a means of rapidly evaluating the relevance of texts. A structural thematic summary can be created for Russian or English documents and presented in Russian or in English.

The second section briefly describes main features of the Thesaurus on Sociopolitical Life. The third section is devoted to a description of the main principles of construction of the thematic representation. The fourth section indicates the main steps of automatic construction of the thematic representation. In the fifth section we consider the form of the structural thematic summary. In the sixth section an evaluation of the automatic process of the thematic representation construction is considered.

2. Thesaurus

The Thesaurus on Sociopolitical Life is a hierarchical net of concepts constructed specially as a tool for different applications of automatic text processing. It contains a lot of terms from economical, financial, political, military, social, legislative and cultural spheres.

The Thesaurus has the following main features (for a detailed description see Loukachevitch et al 1999):

- concepts in the Thesaurus have rich synonymic rows including not only nouns and noun groups but also adjectives and verbs, which allow us to recognise the concepts in various text. The number of synonyms of a concept can be up 20 elements, for example for the Russian concept *nature protection*. Considerable efforts were made to find new synonyms of the Thesaurus concepts in texts;

- such extended rows of synonyms necessarily include ambiguous terms. Additional efforts were made to find unambiguous multiword terms with an ambiguous term as a part and to include the multiword terms in the Thesaurus as new concepts or as synonyms to existing concepts;

- conceptual relations used in conventional information-retrieval thesauri (LIV 1984; Subject Headings 1991; UNBIS Thesaurus 1976) such as Broader Term (BT) -- Narrower Term (NT), Related Term (RT) were added by WHOLE -- PART relations and special modifiers to provide for navigation in the Thesaurus net for different goals. Conceptual relations in the Thesaurus are used for automatic query expansion, for recognition of the lexical cohesion in a text, as a basis for detecting the main theme and subthemes of a text, and for term disambiguation;

- the Thesaurus is constantly tested and corrected during automatic text processing;

- every concept of the Thesaurus has English translations and row of English synonyms (30 thousand entries).

The Sociopolitical thesaurus differs from conventional information-retrieval thesauri (LIV 1984; Subject Headings 1991; UNBIS Thesaurus 1976) and from such linguistic resources as WordNet (Miller et al 1990) and EuroWordNet (Climent et al 1996).

The goal in developing a conventional information retrieval thesaurus is to describe terms necessary for representation of main topics of documents. More specific terms are not included. Ambiguous terms are provided with scope notes and comments convenient for human subjects. In fact a conventional information retrieval thesaurus describes an artificial language based on a real language of a domain. Human subjects have to use their domain, common sense, and grammatical knowledge not described in a thesaurus in order to index documents. Therefore conventional information-retrieval thesauri created for manual indexing are hard to utilise in an automatic indexing environment (Salton 1989). To be effective in automatic text processing a thesaurus needs to include a lot of information that is usually missed in thesauri for manual indexing.

On the other hand the Sociopolitical thesaurus differs from such linguistic resources as WordNet (Miller et al 1990) and EuroWordNet (Climent et al 1996):

- it describes terms of the specific domain and does not include words and terms of common language that can be used in texts of any domains;

- all conceptual relations in the Sociopolitical thesaurus are tested from the point of view of the information-retrieval task;

- ambiguity of thesaurus terms is considered relative to the domain: ambiguous terms that almost always have one meaning in the domain are described as unambiguous. Close senses of terms can be clustered to a single concept (Pustejovsky 1995); there are special methods for descsribing relations of clustered senses;

- descriptions of terms include much encyclopaedic information: possible situations, reasons, results, participants, properties and so on, that may be useful for the information retrieval task.

3. The Thesaurus as a linguistic resource for the construction of the thematic representation of a text

The quality of such automatic text processing procedures as text indexing, text categorisation and text summarisation depends on the quality of recognition of the main theme and subthemes of a text. All the procedures should be based on the same document representation in terms of themes and subthemes and the same linguistic resource.

Construction of thematic representation of a text is based on such a property of a connected text as lexical cohesion.

One of main properties of a connected text is cohesion (Halliday and Hasan 1976). Cohesion involves relations between words that connect different parts of the text. Lexical cohesion is the most frequent type of cohesion. It can be expressed by repetitions, synonyms and hyponyms or by words connected with other semantic relations such as whole - part, situation - participant, object - property and so on.

For example, in a fragment of a text FB6-F001-0015 (see Appendix) from the Text Retrieval Conference text collection (Vorhees and Harman 1995) the terms ***border troops*** and

serviceman establish one chain of lexical cohesion in the text; the terms ***poaching***, ***illegal activity*** and ***law*** form other chain of lexical cohesion relations. Such relations can connect sentences of a text without visible markers:

> *The **border troops** 'are not saber rattling' in Russian territorial waters in the Far East as the mass media, especially the Japanese mass media, are attempting to portray it. **Servicemen** have been legally granted the right to utilize all of the tools at their disposal, including weapons, to put a stop to **poaching**. Russian **Border Troops** Commander-in-Chief Colonel-General Andrey Nikolayev stated that to an ITAR-TASS correspondent while stressing that his subordinates are conducting a strict policy to put a stop to the **illegal activities** of foreign boats. He noted that the President of Russia supports the position of the **border troops** for the full observance of the **law** in the country's territorial waters.*

Cohesion relations connect not only sentences of a text, but also the main theme and subthemes of a text between each other. Van Dejk and Kintsch (1983) describe the topical structure of a text, the macrostructure, as a hierarchical structure in a sense that the theme of a whole text can be identified and summed up to a single macroproposition. The theme of the text can be usually described in terms of less general themes which in turn can be characterised in terms of even more specific themes, and so on.

This means that a connected text has its main theme and this main theme can be formulated. Formulation of the main theme names the most important concepts of the text and relations between them (here and below when we say 'concepts of a text' we imply concepts, the terms of which were mentioned in a text). We call concepts of the main theme of a text 'macroconcepts'. Subthemes of the text discuss relations between macroconcepts or important aspects of a macroconcept. To refer to the main theme a subtheme has to include a macroconcept or its related concept; in sentences of a text such references look like lexical cohesion relations.

Relationships between concepts in the main theme can be subdivided into two subtypes. Some of them together name an object. For example, concepts from the example text *border troops* and *Russian Federation* name the object *'border troops of the Russian Federation';* concepts *Japan* and *boat* name the object *'Japanese boats'*. Relations between such concepts are considered as known and usually are not discussed in the text. Various combinations of such concepts and their related concepts often occur together in the text to refer to the corresponding objects, as in for example, *'Russian Border Troops Commander-in-Chief Colonel-General Andrey Nikolayev'***.**

Relationships between other macroconcepts are discussed in subthemes of the text: how border troops plan to struggle with poaching; what their needs are in order to overcome the problem; the results; how Japanese fishermen poach; how they interact with border troops, and so on. Therefore combinations of such macroconcepts and their related concepts also often occur together in clauses and sentences of the text.

Taking into account the presentation of macroconcepts and their relations in the text we made three main assumptions:

1) preliminary knowledge, especially knowledge about relations between concepts of a text is one of the most important factors in presentation of the main theme of a text in terms of subthemes;

2) the macrostructure of a text is based on nodes of semantically related concepts (all concepts in a node are semantically related to the centre of the node): concepts related to

macroconcepts are used in subthemes for reference to the main theme, concepts related to subthemes concepts are used in more specific subthemes for reference to subthemes of upper levels;

3) concepts semantically related to macroconcepts and belonging conceptual nodes with a macroconcept as a centre are situated near each other in a text more often than concepts of other nodes.

We can restore the conceptual net of a text using the Thesaurus. For every concept of a text we take direct thesaurus relations with other textual concepts from the Thesaurus or automatically infer them using properties of thesaurus relations. This gives us the conceptual net, or so-called 'thesaurus projection', of the text.

The received conceptual net can be subdivided into conceptual nodes. We call a set of concepts related to the same concept the 'thematic node'. The concept that all concepts of the thematic nodes are related to is called a 'thematic centre'. Thematic nodes with macroconcepts as thematic centres are called the 'main thematic nodes 'of a text. Thematic nodes with concepts of various subthemes of a text as thematic centres are called 'specific thematic nodes'.

It is not necessary to create thematic nodes around every concept of the text. We supposed that the thematic centre had to be more important for text content than other concepts of the thematic node and that it had to be somehow stressed in the text. It can be used in the title or in the beginning of the text or it can have the highest frequency among related concepts. Thematic nodes can be constructed around such concepts.

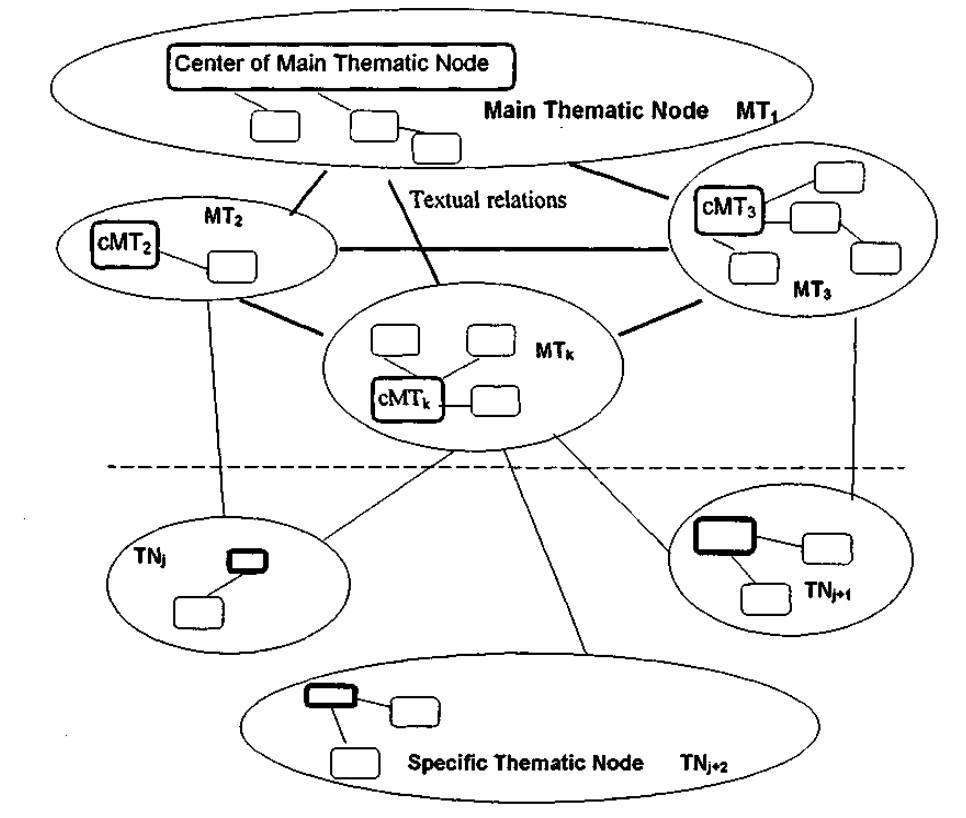

Fig. 1: Hierarchy of thematic representation

To estimate distribution of concepts of different thematic nodes in a text, we use the notion 'textual relation': a given concept has textual relations with those concepts of the text that are located no further than three concepts from the given concept (location order is not important).

Thus the context of a concept occurrence is determined by the quantity of meaningful elements. Other words (not from the Thesaurus) are not included in the count. Textual relations pass through sentence borders and are interrupted only by paragraphs.

The restriction of 'three concepts' of textual relations was derived experimentally. It means that every occurrence of a concept is considered within a set of seven neighbouring concepts.

We incorporate constructed thematic nodes in the'thematic representation' of the text. The thematic representation of text is a hierarchical structure of concepts where concepts semantically or thematically related to thematic centres are gathered in thematic nodes. Thematic nodes whose thematic centres can characterise contents of the text are main thematic nodes. A hierarchy of thematic representation characterises the importance of terms in the text: the thematic centre is more important than other terms of the thematic node; terms of main thematic nodes are more important than the terms of other thematic nodes.

3.1. Related work on lexical cohesion

Recent works research lexical cohesion, expressed mainly by repetitions, synonyms and hyponyms and the construction of 'lexical chains'. A lexical chain is a chain of words in which the criterion for inclusion of a word is some kind of cohesive relationship to a word that is already in the chain (Morris and Hirst 1991). Morris and Hirst also proposed a specification of cohesive relations based on Roget's Thesaurus. Hirst and St-Onge (1997), Barzilay and Elhadad (1997) construct lexical chains based on WordNet relations.

The main stages in the proposed construction of lexical chains are as follows:

- the construction of lexical chains begins from the first words of a text;

- to insert the next word, its relations with members of existing lexical chains are checked. If there are such relations with any element of a chain then the new word is inserted in the chain. Only one lexical chain can be chosen. Otherwise a new lexical chain can be started. The procedure usually includes disambiguation of words.

Construction of lexical chains by Hirst G. and St-Onge D. (1997) and Barzilay R. and Elhadad M. (1997) differs from the construction of our thematic nodes in the following way.

1) They use WordNet as a linguistic resource. However, WordNet was not created as a tool for automatic text processing and it lacks such necessary information such as relations between different parts of speech and such conceptual relations as situation-participant, situations - domain, object - property and others (Climent et al 1996).

In our Thesaurus we tried to describe different types of conceptual relations that can be useful for the detection of lexical cohesion in texts, and we specially tested the Thesaurus as a source of cohesive relations in texts.

2) Deficiency in the scope of conceptual relations causes simplification of the whole picture of used cohesive relations and the structure of lexical chains. For example, a term in a text can belong to more than one lexical chain constructed before it is used. In our approach concepts can belong up to three different thematic nodes.

3) The construction of lexical chains begins from the first sentences of a text and is determined by the development of a text. We tried to reconstruct the overall conceptual basis of a text and made decisions about cohesive relations estimating the whole derived conceptual structure. The thematic node has a definite structure as the central concept and concepts related to the centre, which are interpreted as references and aspects of the central concept.

4) To find the strongest lexical chains in a text Hirst and St-Onge (1997) and Barzilay and Elhadad (1997) use different formula and thresholds. But the proposed formulae and thresholds are very dependent on types and sizes of texts. We found the topological principle of choosing the main thematic nodes that allowed us to process texts of various types and sizes.

5) The techniques of construction of lexical chains were implemented for very restricted tasks and tested on relatively small text collections. Our thematic representation became a basis of various practical applications of automatic text processing of megabytes of Russian and English texts.

4. Construction of the thematic representation

4.1. Identification of terms in texts

Text units are compared with terms of the Thesaurus using morphological representation of the text and terms. If the same fragment of a text corresponds to different concepts of the Thesaurus, ambiguity of the text unit is indicated.

Texts can include names that coincide with terms of the Thesaurus. A name that corresponds to a term of the Thesaurus but has different spelling (capital letters, quotes) is marked as an ambiguous term.

After comparison with the Thesaurus the text is represented as a sequence of concepts. All terms of any concept are represented by the concept and are not differentiated further.

On the basis of the whole set of concepts of the text the thesaurus projection of the text is constructed (see section 3). Figure 2 shows a fragment of the thesaurus projection of the example text.

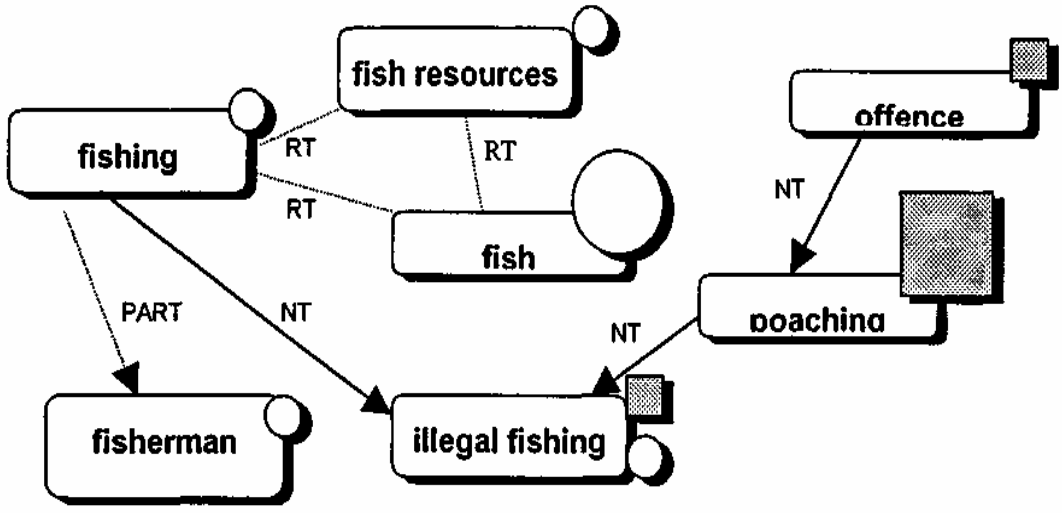

Fig.2. Fragment of Thesaurus projection

4.2. Disambiguation of terms using thesaurus projection

Concepts corresponding to different meanings of ambiguous terms also participate in construction of the thesaurus projection for a text. Using the thesaurus projection a proper meaning of an ambiguous term is chosen.

For every meaning of an ambiguous term the following conditions are verified:

1) A concept corresponding to a meaning of the ambiguous term is used in text in unambiguous form, for example, term *financial capital* is an unambiguous term for the concept *capital(finance)* and *capital* is an ambiguous term for this concept;

2) A concept corresponding to a meaning of the ambiguous term is related to other concepts in the thesaurus projection. For example, concept *public organisation* is connected by relationship HYPONYM with the concept *political party* that corresponds to one of the meanings of ambiguous term *party*.

If one of the conditions is met we consider that the text 'supports' this meaning of the ambiguous term. If the text supports only one meaning of the ambiguous term the corresponding concept is chosen. If the text supports more than one meaning of the term we look through concepts that are the nearest ones to every usage of the ambiguous term and choose the meaning of the concept supported by the nearest concepts.

Only chosen concepts participate in further processing of the text.

4.3. Construction of thematic nodes

The creation of the thematic nodes begins from choosing the thematic centres. At first concepts mentioned in the title and first sentence of the text can gather all related concepts from the thesaurus projection and become the thematic centres of thematic nodes. Then the most frequent concepts of the text can become thematic centres. A concept included in a thematic node cannot become the thematic centre of a new thematic node.

Let us consider document FB6-F001-0015 from TIPSTER Text Collection. Some thematic nodes that were constructed during automatic processing of the example text (the right column represents concept frequency in the text) are as follows:

Russia (Russian)	10
Far East	1
Curile	1
President of Russia	1
state (country)	6
territorial waters	9
ocean	3
water transport (ship, schooner)	11
island	1
state (country)	6
fish	10
fish resources	1

natural resources	3
fishing (catch fish)	5
fishing vessel	3
fisherman	2
illegal fishing	1
poaching (poach)	5
offence (illegal activity, violence)	3
illegal fishing	1

Figure 2 represents two intersecting thematic nodes: a thematic node with the thematic centre *fish* and a thematic node with the thematic centre *poaching.*

4.4. Determination of the status of the thematic node

During the comparison of the text with the Thesaurus terms, textual relations of every concepts (that is neighbour concepts) are collected. As a result we obtain a set of textual relations for every concept of a text. For example, here are fragments of a set of textual relations of concept *fish* received during processing of the text (on the right side frequency of textual relations is indicated):

fish	
Japan	4
territorial waters	3
Russian Federation	2
pouching	2
border guards	1

Textual relations between concepts are determined at the stage of comparison of text with the Thesaurus. After construction of thematic nodes the textual relations of concepts in each thematic node are summed up, and we derive the textual relations between thematic nodes.

Let us consider fragments of textual relations between thematic nodes in the above example. Thematic nodes are represented by their thematic centres; numbers to the right are the total frequency of textual relations between thematic nodes; textual relations are given for the thematic node with thematic centre *fish*.

Fish	
border troops	10
Japan	10
territorial water	9
Russian Federation	7
Poaching	7

In our approach we assume that first of all main thematic nodes are those ones that:

- have textual relations with all other main thematic nodes and

- have a sum of frequencies of textual relations between these nodes that is greater than the sum of frequencies for the same number of other thematic nodes of this text.

In our example the main thematic nodes were thematic nodes with the main concepts *border guards, territorial waters, Russian Federation, fish, poaching, Japan*.

Evaluated in such a way main thematic nodes determine a threshold that distinguishes the main thematic nodes among all the other thematic nodes of a text. This threshold is an average frequency of concepts in determined main thematic nodes. The initial set of main thematic nodes is supplemented with those thematic nodes whose frequency is more than the threshold.

Besides main thematic nodes there are specific thematic nodes and mentioned concepts. Specific thematic nodes represent the primary characteristics of the main topics discussed in the text. Specific nodes are those thematic nodes that have textual relations with at least two different main thematic nodes. Concepts that are not elements of main or specific thematic nodes are called mentioned concepts.

Specific thematic nodes are as follows:

logistics *mass media*

equipment *correspondent*

computer

Mentioned concepts are *legislator, expert, ice situation* **....**

Thus all concepts of the text are divided into five classes of different importance for the text:
- main concepts of main thematic nodes,
- other concepts of main thematic nodes,
- main concepts of specific thematic nodes,
- other concepts of specific thematic nodes,
- mentioned concepts.

5. Structural thematic summary of a text

We made intensive use of the thematic representation to verify the Thesaurus. We used the following procedure: thematic representation of texts were produced. Our specialists compared contents of texts and main thematic nodes identified in the thematic representations of these texts. If they found considerable differences between them, in most cases the reason was some error or inaccuracy in the Thesaurus descriptions. Therefore the described variants of concepts, ambiguity of terms, missed or extra relations between concepts, English translations were verified.

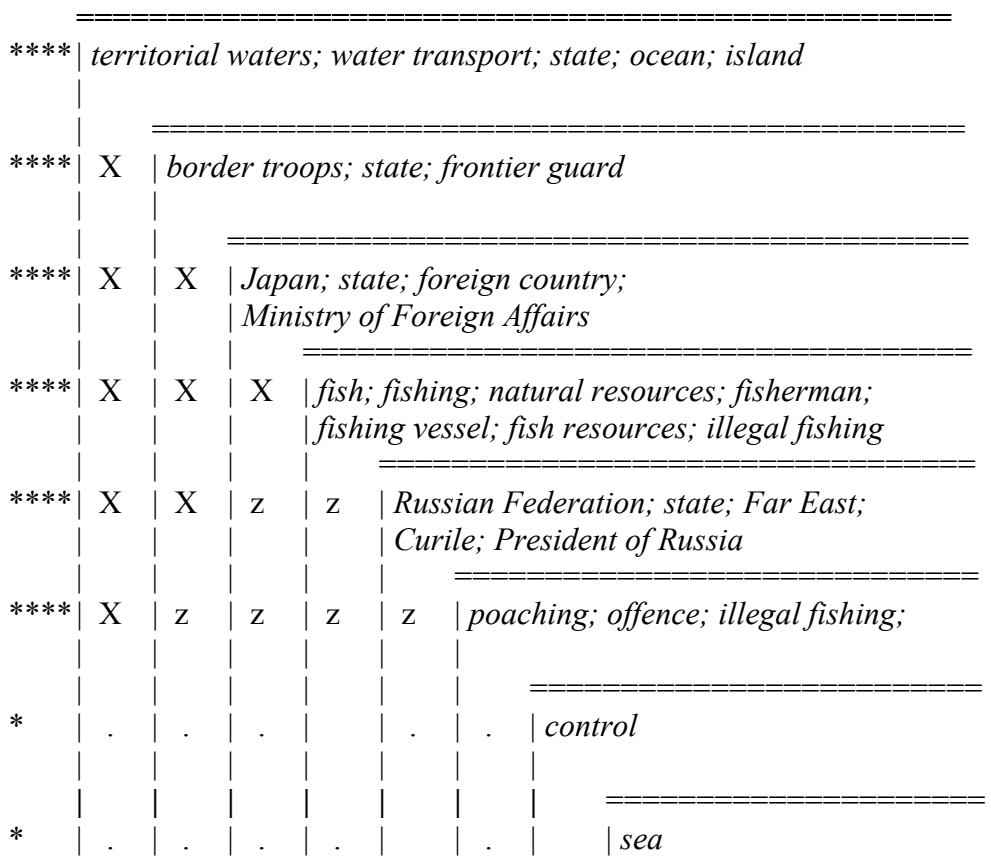

Fig.3. Structural thematic summary of text FB6-F001-0015

The full thematic representation is too detailed to serve for an evaluation of text relevance. The main topics section alone is not enough for an adequate identification of the contents of documents because it does not represent the aspects of topics discussed in a text. Full main thematic nodes are much more informative but in a large text such thematic nodes can be long.. And we created a new structural form of the most important parts of thematic representation - a structural thematic summary (Figure 3). A structural summary allows us to estimate the contents of a document at first sight.

This is the structural summary for the example text (see the fragment in Appendix).

A structural thematic summary contains the following parts:
- the terms of the main thematic nodes ordered by frequency and situated horizontally;
- the marks of relative frequencies of main thematic nodes denoted by different numbers of '*' ('****' - node frequency is more than 75% of maximum node frequency, '***' - node frequency is more than 50 % of maximum; '**' - node frequency is more than 25 % of maximum, '*' in other cases).
- marks of strength of textual relations between different thematic nodes:
-- 'X'-- very frequent textual relations (frequency of textual relations between nodes is more than 75 % of maximum frequency of textual relations);
-- 'z'-- frequent textual relations (frequency is more than 25% of maximum frequency of textual relations);
-- '.' -- weak textual relations.

6. Evaluation

Our experience showed that thematic representation can be produced for texts of any size and for a wide variety of genres and can be used for different applications.

Thematic representations became a basis for information retrieval (Yudina and Dorsey 1995), automatic text categorisation (Loukachevitch 1997) and text summarisation (Loukachevitch 1998) in University Information System RUSSIA (http://www.cir.ru/eng/).

Over 500 Mb of texts including 100 Mb of Russian official documents (Presidential and governmental decrees 1990-2000), 200 Mb of reports by Russian information agencies, and newspaper articles were processed. The smallest texts had a size of 100 Bytes; one of the biggest texts, the Russian Civil Code, was more than 500 Kb.

We also processed texts in English, such as documents of the 104th and 105th Congresses, documents from the routing task of TREC6 and ad hoc tasks of TREC8 (Dobrov, Loukachevitch and Yudina 1997).

In 1998, using our technology of other text summarisation process - automatic sentence-based text summarisation (Loukachevitch 1998), we participated in the SUMMAC conference in a text categorisation task.

Evaluation in this task was as follows. Given a document, which could be a summary or a full-text source, the human subject determines to which single category of six categories (each of which has an associated topic description) the document is relevant (the sixth category being 'none of the above'). Here the evaluation seeks to determine whether a summary is effective in capturing whatever information in the document is needed correctly to categorise the document. Ten topics were chosen, with 100 documents used per topic. These topics were selected such that they could be grouped into two mutually exclusive classes: environment and global economy.

Participants submitted two summaries: a fixed length summary limited to 10% of the character length of the source and a 'best' summary which was not limited in length.

Our summaries of 'best length' had a maximal F-score (SUMMAC Final Report 1998) . The F-score of our 10% summaries was more than medium.

The system extracted sentences in correspondence with the constructed thematic representations. The usage of terms in different main thematic nodes in a single sentence was the main rule used in choosing this sentence for a summary. The constructed set of thematic nodes allowed us to control whether all concepts of the main theme of a text were mentioned within the restricted volume of the summary.

The system was fully based on the Thesaurus knowledge and could not include any processing of manifold proper names, which were very important in the texts. Notwithstanding we received good results. Therefore we consider our results in this competition as confirmation of the quality of our representation of text contents and representation of knowledge.

7. Conclusions

We proposed a new technique for representation of contents of documents. A structural thematic summary presents the main theme and subthemes of a document, which are simulated by sets of semantically related terms. A structural thematic summary can be

constructed for documents of various sizes and genres. It is particularly useful for users of multilingual text retrieval systems.

Clarity and readability of the structural thematic summary is provided by:
- translation of multiword terms, which is easier than translation of a single word or a sentence of a text;
- automatic disambiguation of ambiguous terms;
- the automatic knowledge-based clustering of semantically related terms.

References

Boguraev, B., Kennedy Ch., Bellamy R., Brawer S., Wong Y. and Swartz J. (1998) 'Dynamic Presentation of Document Content for Rapid On-Line Skimming', in *Proceedings of the AAAI'98 Spring Symposium on Intelligent Text Summarisation - AAAI Technical Report SS-98-06*: 118-27

Brazilay R. and Elhadad M. (1997) 'Using Lexical Chains for Text Summarisation', ACL/EACL Workshop on Intelligent Scalable Text Summarisation, Madrid

Climent S., Rodriguez H. and Gonzalo J. (1996) 'Definitions of the links and subsets for nouns of the EuroWordNet project'. Deliverable D005, WP3.1, EuruWordNet, LE2-4003. Also in

Dijk and Kintsch (1983)

Dijk T.A. van and Kintsch W. (1983) *Strategies of Discourse Comprehension*. New York: Academic Press

Dobrov B., Loukachevitch, and N., Yudina T. (1997) 'Conceptual Indexing Using Thematic Representation of Texts', in *The Sixth Text Retrieval Conference (TREC-6),* NIST Special Publication 500-240: 403-12

Fellbaum, C. (ed.) (1997) *WordNet: An electronic lexical database and some of its applications,* Cambridge, MA: MIT Press

Halliday M. and Hasan R. (1976) *Cohesion in English,* London: Longman

Hirst G. and St-Onge D. (1997) 'Lexical Chains as representation of context for the detection and correction malapropisms', in Fellbaum

LIV (Legislative Indexing Vocabulary) (1994) Congressional Research Service. The Library of Congress, 21st Edition

Loukachevitch N. (1997) ' Knowledge Representation for Multilingual Text Categorization', in *Cross-Language Text and Speech Retrieval. - Proceedings of AAAI Spring Symposium on Cross Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05: 133-42

Loukachevitch N. (1998) 'Text Summarization Based on Thematic Representation of Texts. In: Intelligent Text Summarization', in *Proceedings of the AAAI'98 Spring Symposium on Intelligent Text Summarization*, AAAI Technical Report SS-98-06: 77-84

Loukachevitch N.V., Salii A.D. and Dobrov B.V. (1999) 'Thesaurus for Automatic Indexing: Structure, Developement, Use', in Sandrini: 343-55

Loukachevitch, N. and Dobrov B. (2000) 'Modifiers of Conceptual Relations in Thesaurus for Automatic Conceptual Indexing', in *Proceedings of OntoLex-2000*. In press

Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K. (1990) *Five papers on WordNet*, CSL Report 43, Cognitive Science Laboratory, Princeton University

Morris J. and Hirst G. (1991) 'Lexical cohesion computed by thesaural relations as an indicator of the structure of a text', in *Computational Linguistics*, Vol. 17, No. 1: 21-48

Pustejovsky J., (1995) *The Generative Lexicon*, MIT Press

Salton G. (1989) *Automatic Text Processing. The Analysis, Transformation and Retrieval of Information by Computer,* Reading MA: Addison-Wesley

Sandrini P. (ed.) (1999) *Proceedings of Fifth International Congress on Terminology and Knowledge Engineering*, Vienna TermNet

Subject Headings (1991) 14th Edition. Cataloging Distribution Service, Library of Congress, Washington, D.C.

The TIPSTER SUMMAC Text Summarization Evaluation. Final Report 1998. MITRE Technical Report MTR 98W0000138

UNBIS Thesaurus (1976), English Edition, Dag Hammarskjold Library of United Nations, New York

Yudina T. and Dorsey P. (1995) IS RUSSIA: An Artificial Intelligence-Based Document Retrieval System. In *Oracle Select,* 2(2): 12-17

Appendix.

Fragments of Text FB6-F001-0015

'Border Troops `Putina' Exercise to Control Poaching

The border troops 'are not saber rattling' in Russian territorial waters in the Far East as the mass media, especially the Japanese mass media, are attempting to portray it. Servicemen have been legally granted the right to utilize all of the tools at their disposal, including weapons, to put a stop to poaching. Russian Border Troops Commander-in-Chief Colonel-General Andrey Nikolayev stated that to an ITAR-TASS correspondent while stressing that his subordinates are conducting a strict policy to put a stop to the illegal activities of foreign boats. He noted that the President of Russia supports the position of the border troops for the full observance of the law in the country's territorial waters.'

...

'Incidentally, Japanese fishermen (read poachers) have learned about the operation beforehand that is called upon to put pressure on them. This is certainly how we can explain why they have recently stepped up their activities. So, just from 26 March through 1 April and only in the Southern Kurile direction (Izmena Strait and Tanfilyev and Anuchin islands), 49 Japanese boats undertook attempts to poach. The schooners penetrated up to 55 cable lengths (one cable length is approximately 200 meters) into Russian territorial waters.'

...

'But then again, the problem is much broader than just putting a stop to poaching in our territorial waters. It is also whether or not we, having established monitoring of fishing, will be able to conduct our own fishing for fish and crabs in these waters using our own men and equipment, in these waters that have been designated by and that are so familiar for Japanese fishermen? That is not an idle question. The proposals to sell fish to the Japanese that are being increasingly loudly stated today are certainly well thought out. How? Quite

legally: By increasing their quota to catch fish in Russian territorial waters for hard currency. In a word, even in this case the border guards will not be standing idly by: Along with fish conservation personnel, they could carry out monitoring and continue to defend Russia's economic interests in the region.'

interNOSTRUM: A Spanish-Catalan Machine Translation System

by

Raül Canals, Anna Esteve, Alicia Garrido, M. Isabel Guardiola,
Amaia Iturraspe-Bellver, Sandra Montserrat, Pedro Pérez-Antón,
Sergio Ortiz, Hermínia Pastor, and Mikel L. Forcada,
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain.
E-mail: **mlf@dlsi.ua.es**

Abstract

This paper describes *interNOSTRUM*, a Spanish-Catalan machine translation system currently under development that achieves great speed through the use of finite-state technologies and a reasonable accuracy for this pair of closely-related languages by using a classical low-level approach which could be described as an advanced morphological transfer strategy.

Keywords: Spanish, Catalan, machine translation, finite-state.

1 Introduction

This paper describes a Spanish-Catalan machine translation system, *interNOSTRUM*. The main reason for the demand of translations from Spanish (the official language of Spain) into Catalan is the impulse toward 'linguistic normalization' in the Catalan-speaking regions (ten million inhabitants and about six million speakers) where Catalan was receding and where the language is now co-official. Catalan and Spanish are two closely-related romance languages with a rather limited syntactical divergence. The *interNOSTRUM* system is currently under development and a prototype has just started to serve the Universitat d'Alacant, a medium-sized university, and the Caja de Ahorros del Mediterráneo, one of the largest savings banks in Spain. These two institutions started and currently fund this three-year project (1999-2001), which has a staff of two linguists and three computer engineers. Even though translation accuracy and vocabulary coverage can still be much improved, the speed of the system (thousands of words per second or millions of words per day on a 1999-model desktop machine acting as an Internet server) has prompted its use as a system to obtain instantaneous rough translations that are relatively easy to turn into publishable documents. These speeds are achieved through the use of finite-state technolog (Roches and Schabes 1997) in most of its modules.

2 Current prototype and future versions

As has been said, even though *interNOSTRUM* is not a finished product yet, it can nevertheless be used to obtain instantaneous rough translations ready for post-edition. Indeed, two of the basic objectives of our project have been, first, to generate an operational version of *interNOSTRUM* as soon as possible (launched November 1999) and, second, to make the

latest stable version available as soon as it is ready. These are the main reasons for its current configuration as a single Internet server.

Currently, *interNOSTRUM* only translates unformatted ANSI or ASCII texts from Castillian Spanish to the central or Barcelona variety of Catalan (the version generating a València variety and a Balearic Island variety will be ready by the end of the project), but both an RTF (Microsoft's Rich Text Format) and an HTML (HyperText Markup Language) versions are about to be launched. We expect to release the inverse (Catalan-Spanish) translator by November 2000.

2.1 Platform

interNOSTRUM currently runs on Linux and may be accessed through an Internet server (www.internostrum.com or www.torsimany.ua.es). It consists of 6 modules that run in parallel and communicate through text channels. Each module is automatically generated from the corresponding linguistic data using compilers written with the aid of yacc and lex, which are standard in Unix environments. The current speed of the system is in the order of 1,000 wps (words per second) on a standard 1999 desktop PC (a 400 MHz Pentium PC).

2.2 Machine translation strategy

interNOSTRUM is a classical indirect machine translation system using an advanced morphological transfer strategy (similar to a *transformer architecture*; Bouillon and Clas 1993) or *direct system* (Arnold 1993) analogous similar to the one used in commercial PC-based machine translation systems. *interNOSTRUM* has six modules (see figure 1): two analysis modules (morphological analyser and part-of-speech tagger), two transfer modules (bilingual dictionary module and pattern processing module) and two generation modules (morphological generator and postgenerator). The six modules automatically generated from data (see table 1).

2.2.1 Modules based on finite-state technology

Four of the modules in *interNOSTRUM*, namely, the *morphological analyser*, the *bilingual dictionary module*, the *morphological generator*, and the *postgenerator* are based on *finite-state transducers* (FSTs) (Roche and Schabes 1997). This allows for processing speeds on the order of 10,000 words per second, which are practically independent of the size of the dictionaries. Another interesting feature of FST-based modules is that they may be made very compact using standard minimization techniques. FSTs read their input symbol by symbol; each time a symbol is read, they move to a new state, and write, also symbol by symbol, one or more output symbols.

The morphological analyser, which is automatically generated (Garrido et al 1999) from a *morphological dictionary* (MD) for the source language (SL). The MD contains the lemmas (canonical or base forms for inflected words), the inflection paradigms, and their mutual relationships. The subprogram reads the text or *surface* forms (SF) and writes, for each surface form, one or more *lexical forms* (LF) consisting of a lemma, a part of speech, and inflection information.

The bilingual dictionary module is called by the pattern processing module (see below); it is automatically generated from a file that contains the bilingual correspondences. The program reads a SL LF and writes the corresponding target-language (TL) LF.

The morphological generator performs basically the reverse of morphological analysis, but applied to the TL. The morphological generator is generated from a MD for the TL.

The postgenerator: Those SF involved in apostrophation and hyphenation (such as clitic pronouns, articles, some prepositions, etc.) activate this module which is otherwise asleep. The postgenerator is generated from a file containing the corresponding rules for the TL.

The division of a text in words has some nontrivial aspects. On the one hand, there are a number of word groups that cannot be translated word for word and may be treated as fixed-length *multiword units* (MWU); they are gradually being incorporated to the bilingual dictionary. Examples: Sp. *con cargo a* ↔ Cat. *a càrrec de* ('at the expense of'); Sp. *por adelantado* ↔ Cat. *per endavant* ('in advance'); Sp. ***echar*** *de menos* ↔ Cat. ***trobar*** *a faltar* ('to miss [someone]'); in the last example, the MWU has a variable element that may be inflected (in boldface); MWUs with inflection have just started to be incorporated into *interNOSTRUM*'s dictionaries. On the other hand, combinations of certain verb forms and enclitic pronouns are written in Spanish as a single word; these combinations occur with ortographical transformations such as accent marks or loss of consonants: Sp. *dámelo = da + me + lo* ↔ Cat. *dóna + me + lo = dóna-me'l* ('give it to me!'); Sp. *presentémonos = presentemos + nos* ↔ Cat. *presentem + nos = presentem-nos* ('let us introduce ourselves').

Figure 1: Basic *interNOSTRUM* modules

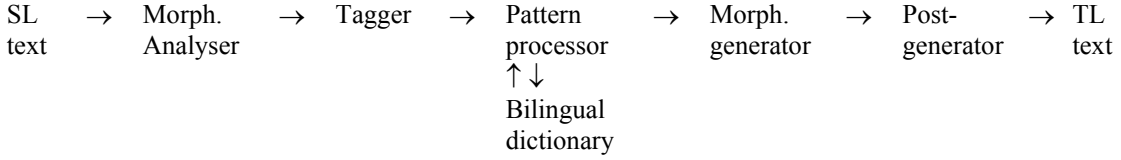

Table 1: Automatic generation of *interNOSTRUM*'s modules from linguistic data

LANG-UAGE	LINGUISTIC DATA	GENERATION PROGRAM	INTERNOSTRUM MODULE
SL	morphological dictionary	Morphological analyser compiler	morphological analyser
SL	Morphologically analysed corpus	Tagger trainer	tagger
SL, TL	bilingual dictionary	Bilingual dictionary compiler	bilingual dictionary module
SL, TL	pattern processing rules	Pattern processing rule compiler	pattern processing module
TL	morphological dictionary	Morphological generator compiler	morphological generator
TL	apostrophe & hyphen rules	Postgenerator compiler	postgenerator

2.2.2 The part-of-speech tagger

Most lexical ambiguities fall into two main groups: *homography* (when a SF has more than one LF or analysis) and *polysemy* (when the SF has a single LF but the lemma may have more than one interpretation).

The lexical disambiguation module or *part-of-speech tagger* uses a language model based on trigrams (sequences of three lexical categories) to solve those homographs ocurring in Spanish texts that present a category ambiguity. The model's parameters reflect the frequencies observed for each trigram in a reference text corpus; the tagger assigns a probability to each possible disambiguation of a sentence containing a lexical categorial

ambiguity and the most likely disambiguation is chosen. We are currently fine-tuning the tagset used and building a larger training corpus to improve the performance of this module and to address homographies inside the same lexical category. Polysemic words will be avoided through the use of a *controlled Spanish* biased toward banking and administration applications (see section 3).

2.2.3 The pattern processing module

In spite of the great similarity between Spanish and Catalan, there are still a number of important grammatical divergences: modal constructions --Sp. *tienen que firmar* ↔ Cat. *han de signar* (``they have to sign")--; gender and number divergences --Sp. *la deuda contraída* (fem.) ↔ Cat. *el deute contret* (masc.; ``the assumed debt")--; dropping of prepositions before *que* --Sp. *la intención de que el cliente esté satisfecho* ↔ Cat. *la intenció que el client estigui satisfet* (Engl. ``the intention that the customer be satisfied")--; relative constructions using *cuyo* (``whose"), absent in Catalan --Sp. *la cuenta cuyo titular es el asegurado* ↔ Catalan *el compte el titular del qual és l'assegurat* (Engl. ``the account whose owner is the insured person").

These divergences have to be treated using suitable grammatical rules. *interNOSTRUM* uses a solution which may also be found in commercial MT systems. It is based on the detection and treatment of predefined sequences of lexical categories (*patterns*) which may be seen as rudimentary phrase-structure constructs: for example, **art.-noun** or **art.-noun-adj.** are two possible valid noun phrases. Those sequences known to the program constitute its pattern *catalog*. This module works as follows:

- The text (morphologically analysed and disambiguated) is read left to right, one LF at a time.
- The module searches, starting at the current position in the sentence, for the longest LF sequence that matches a pattern in its pattern catalog (for example, if the text starting in the current position is ``una señal inequívoca..." (``an unmistakable signal"), it will choose **art.-noun-adj.** instead of **art.-noun**).
- The module operates on this pattern (to propagate gender and number agreement, to reorder it, to make lexical changes) following the rules associated to the pattern.
- Then, the pattern processing module continues immediately after the pattern just processed (it does not visit again any of the LFs on which it has operated).

When no pattern is detected in the current position, the program translates one LF literally and restarts at the following LF. ``Long-range" phenomena such as subject-verb agreement require the propagation of information from one pattern to the following ones; we are currently working on this aspect.

The pattern processing module is automatically generated from a file containing rules that specify the patterns and the associated actions. This is the slowest module (around 1,000 wps), compared to the 10,000 wps of the rest of the modules. The current catalog only contains a few patterns.

3. A translation example

To convey a rough idea of the performance of *interNOSTRUM* in its current state of development (August 2000 version) a short example (randomly chosen from an Internet

newspaper) is translated from Spanish to Catalan and the minimal editing operations to render the translation acceptable are shown in the translation. Most of the observed problems are due to the lack of coverage of current dictionaries, which are expected to be reasonably complete at the end of the project.

Spanish text: Fujimori deja el poder y convoca elecciones generales a las que no se presentará. Lima. -- El presidente de Perú, Alberto Fujimori, ha anunciado por sorpresa en un mensaje televisado a la nación la convocatoria de nuevos comicios y ha precisado que 'en esas elecciones generales, de más está decirlo, no participará quien habla'. Tras 10 años en el poder, el gobernante peruano, el más veterano en Latinoamérica después de su colega cubano, Fidel Castro, presentó así lo que ha sido interpretado como una dimisión en toda regla. La presunta implicación de Vladimiro Montesinos, el más íntimo colaborador de Fujimori, en diferentes actos de corrupción y atentados desde el poder contra el Estado de derecho es la causa directa de la caída del gobernante peruano.

Catalan text with corrections: Fujimori deixa el poder i convoca eleccions generals a *les que[correct:* les quals*/* no es presentarà. *Llima [correct:* Lima*/* -- El president de Perú, Alberto Fujimori, ha anunciat per sorpresa en un missatge **televisado [unknown; correct:* televisat*/* a la nació la convocatòria de nous **comicios* [*unknown; correct:* comicis] i ha precisat que 'en aquestes eleccions generals, *de més està [calque; correct:* no cal*/* dir-*lo [correct:* -ho*/*, no *[insert*: hi*/* participarà qui parla'. Després de 10 anys en el poder, el governant **peruano [unknown; correct*: peruà*/*, el més veterà en Llatinoamèrica després del seu col·lega cubà, Fidel Castro,va presentar així el que ha estat interpretat com una dimissió en tota regla. La pressumpta implicació de Vladimiro Montesinos, el més íntim col·laborador de Fujimori, en diferents actes de corrupció i atemptats des del poder contra l'Estat de dret és la causa directa de la caiguda del governant **peruano [unknown; correct*: peruà*/*.

4 Projected support tools for interNOSTRUM

We are currently working on three support tools: (a) a *style assistant* to help authors of Spanish texts avoid many difficult ambiguities using the syntactical, lexical and style rules specified in a controlled Spanish; (b) a *preedition assistant*, for the manual disambiguation of problematic words and structures, by clicking on them to get a menu of options (helpful when the statistical strategy used by the program is unable to make the right choice); and (c) a *postedition assistant*, in which the author will be able to click on a target-language word when he or she suspects that it is an incorrect translation and will allow him or her to substitute it by an alternative, taking into account the original text.

5 Other Spanish-Catalan MT products

interNOSTRUM is still under development and has not been yet thoroughly compared with its competitors, but they are listed here for completeness; we expect the translation accuracy of *interNOSTRUM* to be comparable to its best competitors toward the end the project, with the added benefit of a speed in the range of 1000 wps. There are currently four more Spanish-Catalan MT products available (the first two may be tested through the Internet):

1. INCYTA's Es-Ca (www.incyta.es) is a syntactical transfer system very much in the spirit of METAL. It runs as a server; customers pay by the word (0.02 euro per word).
2. The newspaper *El Periódico de Catalunya* publishes daily a Spanish version and a cover-to-cover translation to Catalan, using a system developed by SoftLibrary (www.softly.es) which

may be described as a very efficient translation memory which draws from the bilingual corpus of the newspaper.

3. SALT (developed by the Generalitat Valenciana, the government of the autonomous region of València) is available through unofficial channels because it has not been officially published. It runs on Windows at about 10 wps but stops very frequently for human assistance with ambiguous words. It may be classified as a direct system with a plethora of ad-hoc strategies. It generates texts of reasonable quality in the València variety of Catalan.
4. AutoTrad's Ara (www.autotrad.com) is basically a commercial, improved version of SALT which generates the Barcelona variety of Catalan and runs without interruption by accumulating all disambiguation dialogues at the end of the translation.

6 Concluding remarks

We have presented *interNOSTRUM*, a Spanish-Catalan machine translation system currently under development that achieves great speed through the use of finite-state technologies and a reasonable accuracy using an advanced morphological transfer strategy.

References

Roche, E. and Schabes, Y. (eds) (1997) *Finite-State Language Processing,* Cambridge, Mass.: MIT Press: 1997: 1-65

Bouillon, P. and Clas, A. (eds) (1993) *La traductique*, Montréal University Presses

Arnold, D. (1993) 'Sur la conception du transfert', in Bouillon and Clas: 64-76

Hutchins, W. J. and Somers, H. L. (1992) *An Introduction to Machine Translation*, London: Academic Press

Gimènez. M. M. i and Forcada, K. L. (1998) 'Understanding PC-based machine translation systems for evaluation, teaching and reverse engineering: the treatment of noun phrases in Power Translator', in *Machine Translation Review (British Computer Society)*, 7:20-27

Garrido, A., Iturraspe, A, Pastor H., Forcada M.L. and Montserrat, S. (1999) 'A compiler for morphological analysers and generators based on finite-state transducers', in *Procesamiento del Lenguaje Natural*, (25):93-98 1999

State and Role of Machine Translation in India

by

Dr. Sivaji Bandyopadhyay

Computer Science & Engineering Department

Jadavpur University, Calcutta – 700 032, India

E-mail : sivaji_ju@vsnl.com.

In a large multi-lingual society like India, there is a great demand for translation of documents from one language to another. Most of the state governments work in the respective regional languages whereas the Union Government's official documents and reports are in bilingual form (Hindi/English). In order to have a proper communication there is a need to translate these reports and documents in the respective regional languages. With the limitations of human translators most of this information (reports and documents) is missing and not percolating down. A machine assisted translation system or a translator's workstation would increase the efficiency of the human translators.

The Ministry of Information Technology, Government of India, (http://www.mit.gov.in) has identified the following domains for development of domain specific translation systems : government administrative procedures and formats, parliamentary questions and answers, pharmaceutical information, legal terminology and important judgements, and so on. The Ministry initiated the TDIL (Technology Development for Indian languages) project in 1990-91 to support R&D efforts in the area of Information processing in Indian languages covering machine translation among others.

A machine aided translation system (*Anusaaraka*) among Indian languages has been built with funding from TDIL project. The *Anusaaraka* system presents an image of the source text in a language close to the target language. In the image, some construction of the source language (which do not have equivalences in the target language) spill over to the output. *Anusaarakas* has been built for five pairs of languages : Telugu, Kannada, Marathi, Bengali and Punjabi to Hindi. They are available for use through e-mail servers. *Anusaarakas* follows the principle of substitutibility and reversibility of strings produced. This implies preservation of information while going from a source language to a target language. For narrow subject areas, specialized modules can be built by putting subject domain knowledge into the system, which produces good quality grammatical output. However, it should be remembered that such modules will work only in narrow areas, and will sometimes go wrong. In such a situation, *Anusaaraka* output will still remain useful. Work is going on in building an English to Hindi *Anusaaraka* system, which will be a test of building a system between two languages which are far apart. The system so developed will be available as free open-source software under GPL. The work on the *Anusaaraka* project started at the Indian Institute of Technology, Kanpur. It is now being carried out at the Language Technologies Research Center, Indian Institute of Information Technology, Hyderabad (http://www.iiit.net/research/ltrc) with

financial support from Satyam Computers Private Limited. The group at the center is being guided by Prof. Rajeev Sangal and can be contacted at the e-mail address: rambabu@iiit.net.

The Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai is working on *MaTra*, a human-aided transfer-based translation system for English to Hindi. The work is supported under the TDIL project. The domain being explored is news, but the approach is applicable to any domain. The system breaks an English sentence into chunks, analyzes the structure and displays it using an intuitive browser-like representation, which the user can verify and correct, after which the system generates the Hindi. Ambiguities are resolved as they occur by checking with the user. A prototype that can translate simple (single verb group) sentences with occasional human intervention has been developed. A document categorization system has been developed that can automatically classify a news item into one of a set of ten categories, based on a statistical model arrived at after training with a manually created training set. The group is currently developing a practical framework for the syntactic transfer of compound-complex sentences from English to Hindi. More information about the system can obtained from the NCST Website (www.ncst.ernet.in/kbcs/NLP.html) and by e-mail from Durgesh Rao, *MaTra* Coordinator (matra@ncst.ernet.in).

The Machine Aided Translation System *ANGLABHARATI*, for translation from English to Hindi for the specific domain of Public Health Campaign has been developed and is being installed at the user(s) sites for field testing. This technology is proposed to be extended to another domain for translation of Financial and Supplementary Rules of Government of India and related correspondence. The *ANGLABHARATI* project was launched by Professor R. M. K. Sinha at the Indian Institute of Technology, Kanpur in 1991 for machine aided translation from English to Indian languages.

The Multilingual Pocket Translator Design Project was undertaken by the Center for Development of Advanced Computing (CDAC) with an view to foreign travellers visiting India. There is a certain type of fixed requirement when a person is travelling. These requirements are categorized, which can help the foreigner to communicate and prevents the system failing in the initial stages. The same pocket translator is useful when a person moves from one state to another within India. Further details about the Pocket Translator can be obtained from http://vishwabharat.tdil.gov.in/pocket_trans.htm and also from the CDAC website http://www.cdac.org.in. Work is also going on at the CDAC for the development of an English-Hindi machine translation system using Tree Adjoining Grammar.

Work on a knowledge-driven generalized example-based Machine Translation system from English to Indian languages is being carried out in the *ANUBAAD* Project at the Computer Science and Engineering Department, Jadavpur University, Calcutta. It is currently translating short single paragraph news items from English to Bengali. Headlines are translated using knowledge bases and example structures. The sentences in the news body are translated by analysis and synthesis. Semantic categories are associated with words to identify the inflections to be attached in the target language and to identify the context in the sentence. Context identification is also done by using context templates for each word. The example base includes the mapping of grammatical phrases from the source to the target language. The methodologies can be used for developing similar systems for other Indian languages. The work is being carried out as part of a Research Award granted to Dr. Sivaji Bandyopadhyay (e-mail : sivaji_ju@vsnl.com) and by the University Grants Commission (UGC) of the Government of India in 1999 (F. 30-95/98(SA-III)).

Evelyne Viegas (ed.) (1999) *Breadth and Depth of Semantic Lexicons*,
Dordrecht/Boston/London: Kluwer Academic Publishers. Hardback, ISBN 0-7923-6039-7

The volume under review covers the main topics of the workshop on Breadth and Depth of Semantic lexicons, held at the university of California, in conjunction with the meeting on the 34th Annual Meeting of the Association for Computational Linguistics, on June 28, 1996, Santa Cruz, California.

Most of the books about computational lexical semantic lexicons deal with the depth (or content) aspect of lexicons, ignoring the breadth (or coverage) aspect. This book presents a first attempt to address both issues: depth (or content) and breadth (or coverage) of computational semantic lexicons, including lexical and conceptual approaches. Moreover, it addresses issues that have not yet been tackled in implemented systems such as the application time of lexical rules. This book embraces several different fields: Linguistics (theoretical, computational), semantics, pragmatics, psycholinguistics, cognitive science, computer science, artificial intelligence, knowledge representation, statistics, and natural language processing.

This book addresses two main issues of capturing regularities in the lexicon:

1. the treatment of lexical ambiguity;

2. lexical rules as a conceptual tool for controlled proliferation of entries.

Whereas the former has been regarded as a topical issue for quite some time, the latter is only now receiving its due attention. This book addresses both issues with a focus on lexical rules as a regulator of breadth and depth of the lexicons.

The book also investigates whether theoretical accounts of the use of lexical rules are too general and underspecified to support actual processing or not. Practical issues connected with the concrete application of lexical rules are also investigated: when to apply the rules; how the rules influence system design; and how to re-examine and adjust the theoretically posited rules in view of practical needs and evidence.

The book under review comprises four sections: lexical rules and underspecification; breadth of semantic lexicons; depth of semantic lexicons; lexical semantics and pragmatics. All these issues are discussed in fifteen chapters via computational and descriptive case studies addressed within the framework of a theory and an implemented system.

Section 1: lexical rules and underspecification

The chapters of this section discuss the following:

1. The different types of lexical rules an NLP (natural language processing) system could use (B. A. Onyshkevich). There are three sets of parameters a system developer needs to take into consideration when starting building a practical NLP system: the first one addresses the scope of linguistic phenomena covered by the lexical rule (LR) mechanism, including derivational word formation, inflectional morphology, and regular polysemy. The second parameter addresses the application of LRs in a computational system (at lexical acquisition time, lexicon load time, or at run time). The third parameter addresses the triggering and constraining mechanisms of LR applicability needed to avoid over-generation. The

categorisation is also used to overview and classify the approaches to LR in the other papers of this book.

2. Analysis of present studies on the advantages and disadvantages of using lexical rules to extend a lexicon (T. Sehitoglu, C. Bozsahin), outlining a lexical organisation for Turkish that makes use of lexical rules for inflections, derivations and lexical category changes to control the proliferation of lexical entries. Lexical rules handle changes in grammatical roles; enforce type constraints and control the mapping of subcategorisation frames in valency-changing operations. Formally, constraints are enforced via a lexical inheritance hierarchy. The design has been tested as part of an HPSG grammar for Turkish. Concerning the application time of the rules, the authors, based on their experience, reckon that a run-time execution of the rules seems to be a far better alternative than precompilation in view of the intensive use of inflections and derivations in Turkish.

3. B. Gillon's chapter investigates the systematic connection between English mass and count nouns. The author addresses the principal morphological and semantic properties of the mass count distinction; the empirical generalizations pertinent to systematic connection between English mass and count nouns, in terms of lexical rules; and how such rules fit with a syntactic and semantic theory of common English noun phrases.

4. A. Sanfillippo's chapter presents an alternative approach to LRs to treat lexical ambiguity. The author suggests that regularities about sense/usage extensibility should be represented via underspecification in the lexicon. Lexical ambiguity is expressed by associating each ambiguous word form with an Underspecified lexical type which subsumes a class of alternations describing all admissible uses of the word. Appropriate lexical usage can then be assessed deterministically by utilizing syntactic and semantic cues gathered during text processing from the local phrasal context in order to ground underspecified word entries.

Section 2: breadth of semantic lexicons

This section deals with the large-scale acquisition of computational semantic lexicons using lexical rules or concept rules as a conceptual tool to extend a lexicon and existing knowledge sources, such as Word Net, Levin's database of alternation verbs and machine readable dictionaries. Some authors gave preliminary results on the evaluation of their approach with respect to word sense disambiguation (B. J. Dorr and D. Jones). The chapter on deverbal adjectives illustrates how lexical rules can be used to proliferate entries. The authors argue for a large scale approach when using lexical rules (V. Ruskin and S. Nirenburg). The usefulness of lexical rules as in practical system is being questioned, as it seems there are more irregularities than can be managed by their mechanism. (K. J. Burns and A. R. Davis).

Section 3: depth of semantic lexicons

The first two chapters focus on semantic phenomena among adjectives (P. Bouillon), nouns and nominal compounds (M. Johnston and F. Busa). The authors here are more interested in investigating the depth of the phenomena they cover rather than the breadth aspect, although scalability is also addressed by some authors. The last two chapters deal with automating the acquisition of semantic lexicons (J. N. Chen and J. S. Chang) and a corpus of free- responses and lexical semantic techniques (J. Burstein, S. Wolf, C. Lu). These authors give preliminary results on the evaluation of their approach with respect to word sense disambiguation for the former, and information extraction for the latter.

Section 4: lexical semantics and pragmatics

The chapters in this section deal with various perspectives on the role of lexical semantics and pragmatics. The first chapter illustrates the Word Net approach to lexicon organization

comparing its approach to a more syntactic-based approach as described in Levin (1993), showing the absence of an isomorphism between syntax and semantics (C. Fellbaum). The second chapter investigates a new way to get around the incompleteness of computational lexicons by using both lexical rules and conceptual rules to create new lexicon entries and new concepts on the fly (E. Viegas). The third chapter discusses two main approaches to (computational) lexical semantics, Supply-side and demand-side, and primary focusing on theoretical and practical aspects in computational linguistics (S. Nirenburg and V. Ruskin). The last chapter discusses lexical rules and lexical underspecification. The authors argue for pragmatic rules of conversation along with contextual knowledge to treat lexical polysemy (S. Helmreich and D. Farwell).

Nadia Mckendrick

References

Leech, G. (1981) *Semantics,* Cambridge University Press.

Levin, B.(1993) *English Verb Classes and Alternations: A Preliminary Investigation,* University of Chicago Press

Conferences and Workshops

The following is a list of recent (i.e. since the last edition of the MTR) and forthcoming conferences and workshops. Telephone numbers and e-mail addresses are given where known (please check area telephone codes).

12-14 April 2000
RIAO 2000: Sixth Conference on Content-Based Multimedia Information Access
Collège de France, Paris, France
e-mail: cidcol@club-internet.fr
http://host.limsi.fr/RIAO

4 May 2000
Embedded Machine Translation Systems Workshop II
Seattle, Washington, USA
http://lamp.cfar.umd.edu/Embedded_MT_Systems

29 May 2000
LREC 2000: Workshop on the Evaluation of Machine Translation
Athens, Greece
E-mail:LREC2000@ilsp.gr.

2-4 June 2000
NLP 2000: Second International Conference on Natural Language Processing.
Filling the gap between theory and practice
Conference and Cultural Center, University of Patras, Greece
http://www.cti.gr/nlp2000
Tel: +3061 960 383, fax: +3061 997 783, e-mail: pinelop@cti.gr

12-16 June 2000
INLG2000: International Natural Language Generation
Mitzpe Ramon, Israel
Department of Computer Science, Ben Gurion University, PO Box 643, Beer Sheva 84105, Israel

3-14 July 2000
NATO Advanced Study Institute on Language Engineering for Lesser Studied Languages
Bilkent University, Ankara, Turkey
Kemal Oflazer, e-mail: ko@cs.bilkent.edu.tr
http://www.cs.bilkent.edu.tr/~ko/ko.html

15-30 July 2000
TeSTIA: Text and Speech Triggered Information Access
8th European Summer School on Language and Speech Communication
Chios Island, Greece
E-mail: elsnet@let.uu.nl
http://www.ilsp.gr/testia/testia2000.html
26-28 July 2000

IMPACTS
Natural Language Generation between Technology and Applications
Schloss Dagstuhl, Saarland, Germany
http://impacts.dfki.de

COLING2000 in Europe
30 July - 4 August 2000: main conference (Saarbrücken)
29-30 July 2000: tutorials (Nancy)
5-6 August 2000: workshops (Luxemburg)
http://www.coling.org

6 August 2000
COLING 2000: 18th International Conference on Computational Linguistics.
Linguistically Interpreted Corpora (LINC-2000)
Luxemburg, 6 August 2000
http://www.coli.uni-sb.de/linc2000

8-12 August 2000
9th EURALEX International Conference
Stuttgart, Germany
http://www.ims.uni-stuttgart.de/euralex

20-25 August 2000
ECAI 2000: 14th European Conference on Artificial Intelligence
Humboldt University, Berlin, Germany
http://www.ecai2000.hu-berlin.de/

8-10 September 2000
OntoLex 2000: Workshop on Ontologies and Lexical Knowledge Bases
Sozopol, Bulgaria
http://www.sirma.bg/ontolex

13-16 September 2000
TSD 2000: Third International Workshop on Text, Speech and Dialogue
Masaryk University, Brno, Czech Republic
Tel: ++420 5 41 512 359, fax: ++420 5 41 212 568, e-mail: tsd2000@fi.muni.cz
http://www.fi.muni.cz/tsd2000/

20-23 September 2000
AMLAP 2000
Architectures and Mechanisms for Language Processing
Leiden, The Netherlands
E-mail: AMLaP@fsw.leidenuniv.nl
http://www.fsw.leidenuniv.nl/www/w3_func/amlap

3-6 October 2000
ACL 2000: 38th Annual Meeting of the Association for Computational Linguistics
Hong Kong
http://www.cs.ust.hk/acl2000/

7-8 October 2000
EMNLP/VLC-2000: Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora
Hong Kong University of Science and Technology
http://www-csli.stanford.edu/~schuetze/emnlp-vlc2000.html
Hinrich Schuetze, GroupFire (hinrich@groupfire.com)
Keh-Yih Su, Behavior Design Corp. (kysu@bdc.com.tw)

10 October 2000
Machine Translation in Practice: From Old Guard to New Guard
Mision Del Sol, Cuernavaca, Mexico
http://www.isi.edu/natural-language/conferences/amta2000/

10-14 October 2000
AMTA-2000: The Association for Machine Translation in the Americas
Envisioning Machine Translation in the Information Future
Mision del Sol, Cuernavaca, Mexico
http://www.isi.edu/natural-language/conferences/

16-18 October 2000
TALN 2000: Traitement Automatique du Langage Naturel
École Polytechnique Fédérale de Lausanne
http://liawww.epfl.ch/taln2000/ *

19-20 October 2000
ROMAND 2000: First workshop on Robust Methods in Analysis of Natural Language Data
Swiss Federal Institute of Technology, Lausanne
Tel: +41 21 693 52 97, fax: +41 21 693 52 78, e-mail: Vincenzo.Pallotta@epfl.ch
http://lithwww.epfl.ch/romand2000/

3 November 2000
CLIN 2000: The 11th Computational Linguistics in the Netherlands Meeting
Department of Computational Linguistics, Tilburg University, The Netherlands
http://ilk.kub.nl/clin2000

3-4 November 2000
CULT 2000: Second International Conference on Corpus Use and Learning to Translate
Bertinoro, Italy
Tel: +39 0543 450 307/304, fax: +39 0543 450 306, e-mail: mailto:cult2k@sslmit.unibo.it
http://www.sslmit.unibo.it/cult2k/

16-17 November 2000
22nd ASLIB Conference: Translating and the Computer
One Great George Street, London, SW1, UK
Tel: +44 (0) 20 7903 0000, fax: +44 (0) 20 7903 0011, e-mail: nicole.adamides@aslib.co.uk
http://www.aslib.co.uk

16-18 November 2000
DAARRC2000: Discourse, Anaphora and Reference Resolution Conference
Lancaster University, UK
Fax: +44 1524 843085, e-mail: a.mcenery@lancaster.ac.uk

20–22 November 2000
MT 2000: Machine Translation, Multilingualism and the Millennium
University of Exeter, UK
Tel/fax: +44 (0)1392 264296, e-mail: D.R.Lewis@exeter.ac.uk
http://www.bcs.org.uk/siggroup/sg37.htm

23-25 November 2000
TDCNET: Language Technology Conference
Cologne, Germany
http://www.fbi.fh-koeln.de/DEUTERM/ivsw2000E.htm

23-25 November 2000
26th Annual Conference: International Association Language and Business
Language Technologies for Dynamic Business in the Age of the Media
University of Applied Sciences, Cologne
http://www.fbi.fh-koeln.de/DEUTERM/ivsw2000E.htm

10-11 January 2001
Fourth Annual CLUK Research Colloquium
University of Sheffield, UK
http://www.dcs.shef.ac.uk/~njw/CLUK4

10-12 January 2001
IWCS-4: Fourth International Workshop on Computational Semantics
Tilburg, The Netherlands
http://www.sigsem.org/iwcs4.html

March – September 2001
Text Summarization and Document Understanding Conference
The National Institute of Standards and Technology is beginning a new evaluation series in the area of text summarization, tentatively called the Document Understanding Conference DUC). Plans call for the creation of reference data (documents and summaries) for training and testing. The training data will be distributed in March 2001, test data distributed in June, and results due for evaluation 1 August 2001. A workshop will be held in September to discuss these results and to make further plans.
E-mail: marcu@isi.edu.
http://www-nlpir.nist.gov/projects/duc/.

15-17 March 2001
CUNY 2001: 14th Annual Meeting of the CUNY Conference on Human Sentence Processing
Philadelphia, USA
http://www.ircs.upenn.edu/cuny2001

30 March – 2 April 2001

Corpus Linguistics 2001
Lancaster, UK
E-mail: mcenery@comp.lancs.ac.uk

2-7 June 2001
Language Technologies 2001: Second Meeting of the North American Chapter of the Association for Computational Linguistics
Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
http://www.cs.cmu.edu/~ref/naacl2001.html

6-11 July 2001
ACL-2001
Toulouse, France
http://www.irit.fr/ACTIVITES/EQ_ILPL/aclWeb/acl2001.html

14-16 June 2001
BI-DIALOG 2001: Fifth Workshop on the Semantics and Pragmatics of Dialogue
Bielefeld University, Germany
http://www.uni-bielefeld.de/bidialog/

18-20 June 2001
IcoS-3: Inference in Computational Semantics
Siena, Italy
http://www.cs.cmu.edu/~kohlhase/event/icos3/

27–29 June 2001
LACL 2001: Fourth International Conference on Logical Aspects of Computational Linguistics
Le Croisic, France
http://www.irisa.fr/LACL2001

8-10 August 2001
The 2001 Asian Association for Lexicography (ASIALEX) Biennial Conference
Yonsei University, Seoul, Korea
E-mail: asialex@lex.yonsei.ac.kr

20-24 August 2001
ESSLLI Workshop on Information Structure, Discourse Structure and Discourse Semantics
Helsinki, Finland
http://www.helsinki.fi/esslli

September 2001
Eighth International Conference on Translation
Langkawi, Malaysia
http://web.uum.edu.my/ict/

5-7 September 2001
RANLP-2001: Recent Advances in Natural Language Processing
Tzigov Chark, Bulgaria

http://www.lml.bas.bg/ranlp2001

7-9 September 2001
PALC 2001: Practical Applications in Language Corpora
Department of English Language, Lodz University
Tel: ++48 42 639 02 20 20, fax: ++48 42 639 02 18 20
e-mail: corpora@krysia.uni.lodz.pl

11-14 September 2001
PACLING 2001
Kitakyushu International Conference Center, Kitakyushu, Japan

18-22 September 2001
MT Summit VIII
Santiago de Compostela, Galicia, Spain
http://www.eamt.org/summitVIII/

MEMBERSHIP: CHANGE OF ADDRESS

If you change your address, please advise us on this form, or a copy, and send it to the following (this form can also be used to join the Group):

Mr. J.D.Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks, Kent TN13 1QU
U.K.

Date: ...../...../.....

Name: .....
Address: .....
.....
Postal Code: ..... Country: .....
E-mail: ..... Tel.No: .....
Fax.No: .....

Note for non-members of the BCS: your name and address will be recorded on the central computer records of the British Computer Society.

Questionnaire

We would like to know more about you and your interests and would be pleased if you would complete as much of the following questionnaire as you wish (please delete any unwanted words).

1. a. I am mainly interested in the computing/linguistic/user/all aspects of MT.
   b. What is/was your professional subject? .....
   c. What is your native language? .....
   d. What other languages are you interested in? .....
   e. Which computer languages (if any) have you used? .....

2. What information in this Review or any previous Review have you found:
   a. interesting? Date .....
      .....
      .....
   b. useful (i.e. some action was taken on it)? Date .....
      .....
      .....

3. Is there anything else you would like to hear about or think we should publish in the *MT Review*?
   .....
   .....
   .....
   .....

4. Would you be interested in contributing to the Group by,

   a. Reviewing MT books and/or MT/multilingual software
   b. Researching/listing/reviewing public domain MT and MNLP software .....
   c. Designing/writing/reviewing MT/MNLP application software .....
   d. Designing/writing/reviewing general purpose (non-application specific) MNLP .....
      procedures/functions for use in MT and MNLP programming .....
   e. Any other suggestions? .....
      .....
      .....
      .....

Thank you for your time and assistance.