WHY HUMAN TRANSLATORS STILL SLEEP IN PEACE? (FOUR ENGINEERING AND LINGUISTIC GAPS IN NLP)

Paola Velardi

Istituto d'Informatica, via Brecce Bianche, Ancona, Italy

ABSTRACT

Because they will keep their job quite for a few.

This paper has been inspired by a recent editorial on the Financial Times, that gives a discouraging overview of commercial natural language processing systems ('the computer that can sustain a natural language conversation... is unlikely to exist for several decades'). Computational linguists are not so much concerned with applications but computer scientists have the ultimate objective to build systems that can 'increase the acceptability of computers in everyday situations.' Eventually, linguists as well would profit by a significant break-through in natural language processing.

This paper is a brief dissertation on four engineering and linguistic issues we believe critical for a more striking success of NLP: extensive acquisition of the semantic lexicon, formal performance evaluation methods to evaluate systems, development of shell systems for rapid prototyping and customization, and finally a more linguistically motivated approach to word categorization.

THE ENTANGLED FOREST

In the last decade, formal methods to express syntactic and semantic knowledge (whether in an integrated fashion or not), proliferated to form an entangled forest. New comers seem to prefer inventing a brand-new method, or at least a brand-new name, rather than trying to make sense of the dozens of *-PSG, *-unification-G, *-systemic-G, etc. Semantic languages are relatively fewer, but even fewer are the commonly agreed principia about the type and quality of language phenomena to be expressed.

Different are also the perspectives under which linguists and computer scientists proceed in their work:

Linguists and psychologists are concerned with the nature of human communication, and use the computer as a tool to model very specific, and yet meaningful aspects of language. To them, any phenomenon is worth to bee looked at, no matter how frequent, because the focus in on humans, not on computes.

Computer scientists are interested in building computer programs that can ultimately be useful in some relevant field of social life, as machine translation, information retrieval, tutoring, etc. In order for a NLP system to be successful, it must cover the majority of language phenomena that are prominent to a given application. Coverage here is a primary demand, because the focus is on the use of computers, not on the modeling of mind.

I believe that failing to state clearly these differences has been a source of misunderstanding and scarce cooperation. Recently Jacobs pointed out (Jacobs 1989) that linguists measure the power of a parser against pathological cases, and this very fact 'has been damaging to natural language processing as a field'. Linguists may as well complain that the proliferation of NLP papers listing in detail the computational features of 'THE SYSTEM X' and claiming some 5% better performances, has been damaging to computational linguistics as a field.

The author of this paper does not consider her past (and current) work untouched by these criticisms, but wishes that some more explicit and general re-thinking be shared by the computational linguistics + natural language processing community. This paper was inspired by a recent editorial on the Financial Times (Cookson 1989) that presents an overview of commercial and research systems based on NLP technology. The panorama of commercial systems is quite discouraging: the editorial is spread with such sentences as 'not yet robust enough' 'their grammatical coverage is modest' 'no computer has the background knowledge to resolve enough ..

linguistic ambiguities' and concludes: 'the computer that can sustain a natural free-flowing conversation on a subject of your choice is unlikely to exist for several decades.' On the other side, the author highlights several times the importance of this discipline and its possible applications. He also quotes the UK bank's innovation manager David Barrow who says 'Natural language processing will be a key technology in increasing the acceptability of computers in everyday situation'.

Yet, natural language processing began to appear as a discipline since 1950. Progress has been certainly made, but it is not a striking one, with respect to other disciplines equally mature. Why is that? The reader of this paper should be aware by now he run across one of those where-are-we-now-and-where-are-we-going kind of papers; but we hope he will keep following us in a brief walk through the rough pathway of NLP. But please remember... some (not all) viewpoints expressed hereafter would seem narrow-minded if applied to computational linguistics, but are perfectly reasonable if the perspective is robust NLP.

In my view, the major obstacle to a wider adoption of NLP systems is identified by four engineering and linguistic 'gaps'. Engineering gaps are:

1. Lack of formal evaluation methods (Section 1);
2. Lack of tools and engineering techniques for rapid prototyping and testing of NLP modules (Section 4).

Linguistic gaps are:

1. Poor encoding of the semantic lexicon (Section 2);
2. Poorly motivated models of word categorization (Section 3).

This paper has two strictly related guideline ideas, that I would like to state at the beginning:

1. Breadth is more important than depth: In evaluating the pros and cons of linguistic and computer methods for NLP we should always keep in mind their breadth. Methods that cannot be applied extensively and systematically are simply uninteresting. It is perfectly reasonable, and in fact very useful (despite what Hans Karlgren thinks about) to work on sub-languages, provided that the experiments we set to process such domains are reproducible on any other sub-domain. It is perfectly reasonable to define very fine-grained knowledge representation and manipulation frameworks to express deep language phenomena, provided we can demonstrate that such knowledge can be encoded on an extensive basis. As long as the field of linguistic knowledge representation will neglect the related issues of knowledge identification and acquisition, we cannot hope in a breakthrough of NLP.
2. Domain-dependency is not so bad. One of the early errors in AI was the attempt of devising general purpose methods for general purpose problems. Expert systems have been successful but they lie on the other extreme. Current AI research is seeking for a better compromise between generality and knowledge power. Linguistic knowledge is very vast and a full codification is unrealistic for the time being. I believe that a central issue is to accept the unavoidable reality of domain-dependent linguistic knowledge, and seek for generalizable methods to acquire one such knowledge. As discussed in section 3, I also believe that useful linguistic insights can be gathered by the study of language sub-domains.

1. THE 'TRULY VIABLE' APPROACH

Let us maintain our forest-and-path methaphor. Why is it so difficult to get oriented? The cunning reader of technical papers might have noticed a very frequent concluding remark: 'we demonstrated that XYZ is a viable approach to sentence (discourse, anaphora) analysis (generation)'. But what is 'viable'? Other disciplines developed models and experiments to evaluate a system: one could never claim XYZ viable without a good dial of tables, figures and graphs. Why is it so difficult in the field of NLP?

Very few papers have been published on the evaluation of NLP systems. Some well documented report on large NLP projects provides such performance figures as accuracy, intelligibility and quality, however these figures are not uniformly defined and measured. One good example is the Japanese Project (Nagao 1988). The evaluation is performed by humans, applying some scoring to the system output (e.g. translation quality).

Other papers provide a list of language phenomena dealt with by their systems, or an excerpt of sentence types the system is able to process. These results give at best some feeling about the real power of a system, but by no means can be taken as a formal performance measure.

Two papers address the problem of performance evaluation in a systematic way: (Guida 1896) and (Read 1988). The approaches are rather different: Guida and Mauri attempt an application of standard performance evaluation methods to the NLP discipline, introducing a formal expression for the performance measure of a NLP system. This is an hard task, as it comes out of the last section of the paper, where the formula is applied to a simple system. Nevertheless, we believe this work being seminal: formal methods are the most suitable for an uniform evaluation of NLP systems.

In (Read 1988) a 'sourcebook approach' is pursued. The authors propose a fine-grained cataloguing of language phenomena, to be used as a reference for the evaluation of NLP systems. This method in our view is not in contrast with, but rather complementary to, a formal evaluation. However, the final results of this research are not readily available as yet. A second remark is that in measuring the competence of a system, linguistic issues should be weighed by the 'importance' they have in a given application. It is unrealistic to pretend that a system can address every possible phenomenon, but it must be able to address those phenomena that are prominent to the application domain.

One interesting question is: How do we evaluate the linguistic closure of a sub-language? Here is a list of measures, that have the interesting (to me) feature of being acquirable with the use of computers:

1. Identification of the sub-language by a plot of different root-form types per corpus size;
2. Identification of contexts, by and analysis of word co-occurrences, and identification of semantic relations, by an analysis of functional words;
3. Measures of complexity, to predict the computational tractability of a corpus. Some of these measures are listed in (Kittredge 1987), e.g. presence of copula, conjunctions, quantifiers, long nominal compounds, etc. Others are suggested in the very interesting studies on readability, originated by (Flesh 1946). To our knowledge these methods have never been applied to the study of linguistic closure in NLP, even though they reached a remarkable precision at measuring the effect of sentence structures and choice of words on language comprehension by humans (and consequently by computers).

2. THE WORLD IN A BOX

Language resides in the lexicon: word knowledge is world knowledge. One of the major limitation of current NLP systems is a poor encoding of lexical semantic knowledge: the world fits a small box.

The problem with lexica is twofold: First, there is no shared agreement about the type and quality of phenomena to be described in a lexicon. In (Evens 1988) three major competing approaches to meaning representation in lexica are listed: relational semantics, structural semantics and componential/feature analysis. In (Leech 1981) 7 types of meaning are distinguished.

Relational semantics, but for the type and number of conceptual relations (or cases) to be used, shows some uniformity among its supporters for what concerns the structure of the lexicon and the way this information is used to perform semantic analysis. The other approaches highlight much deeper phenomena than the semantic relations between the words in a sentence, but it is a hard task to induce from the literature any firm principle or shared agreement on the type of information to be represented.

In (Velardi forthcoming) it is attempted a more detailed cataloguing of meaning types as found in NLP literature. It is shown that all types of semantic knowledge are in principle useful for the purpose of language understanding applications, but cannot be acquired on an extensive basis because the primary source of such knowledge are linguists and psycholonguistic experiments. Again, relational semantics is somehow more intuitive than other methods and it is easier to acquire, because it can be induced using the evidence provided by texts rather than deduced by pre-defined conceptual primitives. But even then, acquiring more than a few hundred word definitions became a prohibitive task because of consistency, completeness, and boredom problems.

Some work on computer aided acquisition of lexica recently started (Calzolari 1988) (Velardi 1989a,b) (Zernik 1989a) (Jacobs 1988) (Binot 1987); during IJCAI 1989, a workshop was held on this topic (Zernik 1989b). All the above works use corpora or on-line dictionaries as a source of semantic learning, but the methodologies employed to manipulate this texts are very different and still inadequate to the task. Personally, we believe corpora a more adequate source of information than dictionaries.

- on-line dictionaries are not easily available to the scientific community;
- dictionaries mostly include taxonomic

information, that is hardly extracted because of circularity and consistency problems, and because there is no clear method to extract and describe multiple senses in absence of examples;

- the information is not uniform within a given dictionary, and may be very different from dictionary to dictionary depending upon their purpose (e.g. etymological dictionaries, style dictionaries, etc.).
- most of all, the information in dictionaries is very general, whereas in NLP often are required domain-specific categories and definitions.

Corpora provide rich examples of word uses, including idioms and metonymies. It is possible to identify different senses of a word by a context analysis (Velardi 1989a) (Jacobs 1988). In addition, if the corpus used for lexical acquisition is the application domain, one can derive a catalogue of relevant language issues.

In any case, both research on corpora and dictionaries is very promising, and hopefully will provide in the near future more insight and experimental support to meaning theories.

3. THE "IS_A" DILEMMA

The core of any meaning representation method is a conceptual hierarchy, the IS_A hierarchy. People that have experience on this, know how much time-consuming, and unrewarding, is the task of arranging words in a plausible hierarchy. The more concepts you put in, the more entangled becomes the hierarchy, and nobody is never fully satisfied. In (Niremburg 1987) a system is presented to assist humans in entering and maintaining the consistency of a type hierarchy. But this does not alleviate the inherent complexity of grouping concepts in classes.

One could maintain that type hierarchies in NLP systems should not mimic human conceptual primitives, but rather they are a computer method to express semantic knowledge in a compact form and simulate some very partial reasoning activity. Even under this conservative perspective, it is quite natural for the human hierarchy builder to try to make sense of his own taxonomic activity (and get confused) rather than stay with what the specific application requires. Why not introducing such categories as MENTAL_ACT and SOCIAL_PHENOMENON even though the texts to be processed only deals with files and disks? Several institutions devoted large efforts towards the definition of IS_A hierarchies for NLP. Some of these hierarchies are claimed 'general-purpose': to me, this claim is a minus, rather than a plus.

NLP systems have been often presented as a model of human activities. Now, our taxonomic activity is precisely one good example of activity that works very differently than in computers. In computers, hierarchies are used to assert that, if X has the feature Y, and Z is-a X, then Z has the feature Y. Things are in the same category iff they have certain properties in common. This is an objectivist view of categorization that has been proved in several studies inadequate to model human behavior. Objectivism has been argued against in experimental studies by psychologists, anthropologists, and linguists. In his beautiful book (Lakoff 1987) Lakoff lists several phenomena relevant to the activity of categorization, like: family resemblance, centrality, generativity, chaining, conceptual and functional embodiment etc. Only the first of these phenomena has to do with the classical theory of property inheritance. But Lakoff shows that the elements of a category can be related without sharing any common property. The title of his book 'woman, fire and dangerous things' is an examples of apparently unrelated members of a single category in an aborigenal language of Australia. The categorization principle that relates these elements is called by Lakoff the domain-of-experience principle. Woman and fire are associated in myth. Fighting and fighting implements are in the same domain of experience with fire, and hence are in the same class. Birds also are in the same class, because they are believed to be the spirits of dead human-females. Other elements are 'called' in a class by a chaining principle. Element x calls element y that calls z etc.

It is outside the scope of this paper to summarize, or even list, the findings of Lakoff and other researchers on human taxonomic activity. However the literature provides evidence and matter of thoughts concerning the inadequacy of property inheritance as a method to structure linguistic knowledge in NLP systems.

But even if we stay with property inheritance, we should at least abandon the idea of seeking for general purpose taxonomies. Again, corpora are a useful basis to study categorization in sub-worlds. Categories in dictionaries are the result of a conceptualization effort by a linguist. Corpora instead are a 'naive' example of a culturally homogeneous group of people, that draw much unconsciously on their knowledge on the use, and meaning, of words. Corpora are more interesting than dictionaries to study categorization, just like tribes are more

interesting than 'civilized' cultures to anthropologists.

4. GET ACCUSTOMED TO CUSTOMIZATION

The main obstacle to a wider adoption of NLP systems in such activities as information retrieval and automatic translation are reliability and customization. These two issues are clearly related: NLP make errors not because the programs have bugs, but because their knowledge base is very limited. To cope with poor knowledge encoding, ad-hoc techniques are widely adopted, even though the use of ad-hoc techniques is not advertised in papers, for obvious reasons. Ad-hoc techniques are the main cause of long customization time, when switching from one application domain to a slightly different one.

Customization and reliability are in turn related with what we said so far:

- we can't predict the time spent for customization, as it happens in database systems, because methods for knowledge acquisition and knowledge structuring do not exist or are far from being assessed;
- we can't evaluate reliability, because there are not formal evaluation methods for NLP systems.

Again, we came to the same problems. But if we must forcefully abandon the idea of general purpose language processors, at least we should equip ourselves with shell systems and human-computer interfaces that can assist humans in the creation, testing and maintenance of all data-entry activities implied by NLP systems. This paper showed that in semantics there are not as yet assessed theories. In syntax, we have too many, but not systematically tested. Shells and interfaces are useful at:

1. performing a wider experimentation of different theories;
2. making the data-entry activity by humans more constrained or at least supervised;
3. render the customization activity to some extent forecastable;
4. ensure consistency with the linguistic principia embraced by the system designers.

In the field of Expert Systems, shells began to appear when the expert system technology was well assessed. May be shells and interfaces have been disregarded so far by the computational linguistic community because they are felt immature, given the state of art, or just because we are so much affectionate toward the idea of encoding the world.... However, several activities concerned with NLP systems can be computerized or computer-assisted. We already mentioned the work by Niremburg et al. to assist the creation of a concept ontology. A special extension of this system is under experimentation to guide the acquisition of a relational lexicon (Niremburg 1989). Other systems have been presented for prototyping and testing of syntactic parsers (Briscoe 1987) (Bougarev 1988) (Marotta 1990).

5. I DON'T HAVE THE READY RECIPE

I know you knew it! Where-are-we-now papers never offer a panacea. This is a position paper: it did not present solutions, rather it pinpointed to problems and, where available, to current promising research (rather immodestly, some is of our one). The following is a summary list of what the author considers her own guidelines for future work:

- Performance evaluation: never say a method is 'viable' if you can't prove it formally.
- Lexical semantics: don't try to seek for the 'real meaning' of things. Use evidence provided by on-line corpora as a source, and test-bed, of lexical acquisition methods.
- Ontologies: property inheritance is inadequate. Is it possible to implement on a computer some of the observed human mechanisms of categorization?
- Customization: general purpose systems are unrealistic. Build shells and interface systems to allow for a faster and well-assisted customization activity.

ACKNOWLEDGEMENTS

This work has been supported by the European Community under grant PRO-ART 1989

REFERENCES

(Binot 1988) Binot J.l., Jensen K. Dictionary Entries as a source of knowledge for syntactic and other disambiguations Proc. of 2nd Conf on Applied Natural Language Processing Austin, February 1988

(Boguraev 1988) Boguraev B., Carrol J., Briscoe E., Grover C. Software support for practical grammar development in COLING 88 Budapest 1988

(Briscoe 1987) Briscoe E., Grover C., Boguraev B., Carroll J. A formalism and environment for the development of a large grammar of English in IJCAI 1987 Milano, 1987

(Calzolari 1988) N. Calzolari The Dictionary and the thesaurus can be combined in Relational Models of the Lexicon Cambridge University Press, 1988

(Cookson 1989) C. Cookson Why computers need to learn English in Financial Times September 20, 1989

(Evens 1988) M. Evens Introduction in Relational Models of the lexicon M. Evens ed. Cambridge University Press 1988

(Flesh 1946) R. Flesh The Art of Plain Talk in Harper and Brothers 1946

(Guida 1986) Guida G., Mauri G. Evaluation of Natural Language Processing Systems: Issues and Approaches in Proceedings of the IEEE vol.74, n.7 July 1986

(Jacobs 1988) P. Jacobs, U. Zernik Learning Phrases from Texts: A Case Study in AAAI 88 St. Paul, 1988

(Jacobs 1989) P. Jacobs Making sense of lexical acquisition in Proc. of 1st. IJCAI Lexical Acquisition Workshop Detroit 1989

(Lakoff 1987) G. Lakoff Woman, Fire and Dangerous Things: what categories reveal about mind University of Chicago press, 1987

(Leech 1981) G. Leech Semantics Penguin books 1981

(Marotta 1990) Marotta, Pazienza, Pettinelli, Velardi On Parsing, Form-parsing and Meta-parsing submitted, 1990

(Nagao 1988) M.Nagao, Tsujii J., Nakamura J., The Japanese Government Project for Machine Translation in Machine Translation Systems ed. by J.Slocum, Cambridge University Press, 1988

(Niremburg 1987) S. Nirenburg , V. Raskin The subworld concept lexicon and the Lexicon management System in Computational Linguistics n. 13, 1987

(Niremburg 1989) Niremburg S., Raskin V., McCardell R. Ontology based lexical acquisition in Proc. 1st. IJCAI Lexical Acquisition Workshop Detroit, 1987

(Read 1988) Read W., Quilici A., Reeves J., Dyer M., Baker E. Evaluating natural language systems: a sourcebook approach in COLING 88 Budapest 1988

(Velardi 1989a) Velardi, Pazienza Computer aided acquisition of lexical cooccurrences in ACL 89 Vancouver, 1989

(Velardi 1989b) Velardi, Pazienza, Magrini Acquisition of semantic patterns from a natural corpus of texts in ACM-SIGART special issue on knowledge acquisition n. 107 , 1989

(Velardi forthcoming) P. Velardi Acquiring a Semantic Lexicon for Natural Language Processing in U. Zernik ed., Karl Erlbaum assoc., forthcoming

(Zernik 1989) U. Zernik Lexicon Acquisition: Learning from Corpus by Capitalizing on Lexical Categories in IJCAI 1989 Detroit 1989

(Zernik 1989b) U.Zernik ed. First Int. Lexical Acquisition Workshop Proceeding Detroit 1989