Terminology Encoding in View of Multifunctional NLP Resources

Marianna Katsoyannou & Eleni Efthimiou

ILSP-Institute for Language and Speech Processing 6, Artemidos Str. & Epidaurou, 151 25 Paradissos Amaroussiou, Greece {marianna, eleni_e}@ilsp.gr

Abstract

Given the existing standards for organising terminology resources, the main question raised is how to create a DB or assimilated term list with properties allowing for an efficient NLP treatment of input texts. Here, we have dealt with the output of MT and have attempted to improve terminological annotation of the input text, in order to optimize reusability and efficiency of performance. By organizing terms in BD-like tables, which provide various cross-linked indications about head properties, morpho-syntax, derivational morphology and semantic-pragmatic relations between concepts of terms, we have managed to improve functionality of resources and enable better customisation. Moreover, we have tried to view the proposed term DB organisation as part of a global account of the problem of terminology resolution on-processing via grammar based or user-machine interaction techniques for term recognition and disambiguation, since term boundary definition is generally recognised to be a complex and costly enterprise, directly related to the fact that most problem causing terminology items are multi-word units either characterized as fixed or as *ad hoc* or not yet fixed terms.

1. Introduction

Correctly identified terminology crucially increases the reliability of the output of NLP applications which are based on resources such as encoded corpora and machinereadable lexica built according to well-established standardization principles. In this respect, a number of fundamental criteria known to count for the design of text encoding standards, may also provide clues to term recognition and handling in NLP. The issues to be discussed here are based on extensive study of the output of Machine Translation (English to Modern Greek) produced by the EC/SYSTRAN system on a source corpus of medium-large scale, consisting of technical manuals, published by the OECD and dealing with various aspects of R&D surveying. Having focused on systematic errors due to inability to correctly handle a considerable number of terminology items, we have made use of a syntax based multi-word term boundary recognition method (Efthimiou, 1992; Efthimiou & Katsoyannou, 2000) and a user-machine interaction technique for term disambiguation in analysis (Efthimiou & Alexandris, 1998) in order to propose a DB organisation structure, which will allow for optimal management of term related information in MT as well as in other corpus based NLP applications.

2. Problem definition

The need for manual post-editing of the MT output, brought to light a methodological as well as practical issue connected with the problem of terminology resolution: even when marked according to standard corpus encoding principles, terminology-intensive texts still fail to provide "proper/ideal" input for successful parsing.

Apart from difficulty due to the terminology content *per se* (different classes of terminology, text-specific terms loaned from different sublanguages, not yet fixed/*ad hoc* terms, one-to-many translations of a single item, etc.), a major error source was due to the fact that most items falling within the above mentioned categories (mainly

multi-word terms, most often also falling within the *ad hoc* term class) failed to match with proper translations. Thus, indicatively, for a string as [national price indices], the translation output was the pragmatically/semantically wrong string: *[indices of national prices]. Notice that the problem was not related to the rule organisation of the analysis and translation modules but to the unrecognised termhood of the string (for an approach to this problem, see Efthimiou & Katsoyannou, 2000).

Furthermore, a serious source of overgeneration was noticed to be the one-to-many translation problem, exemplified by concepts related to terms with high frequency rates in the relevant texts, for instance, *industry*, *link* or *training*, which find more than one equivalencies in Greek. The same problem has been noticed in the opposite direction as well, e.g. Greek makes no lexical differentiation for concepts as *survey*, *research* and *investigation*. Even worse, the translation(s) of *measure*, *estimate* and *estimation* partly overlap, due to the semantic organization in the lexicon of the specific language.

In recent work (Efthimiou & Katsoyannou, 2000) we have made extensive reference to two major issues in respect to term recognition and disambiguation:

- i) The term boundary definition of multi-word terms by means of the surface syntax of the term string and based on the [+TERM] property of the syntactic head of the construction
- ii) The disambiguation of the one-to-many correspondences by means of a clever user-system interactive mechanism.

These two methodologies were designed to apply onprocessing and had the goal to reduce the analysis problems, due to failure to automatically handle terminology recognition problems reported both in relation to statistical (Church & Hunks, 1990; Frantzi & Ananiadou, 1997) as well as syntactic (Bourigault, 1992) approaches to terminology resolution, or to their combination (Dagan & Church, 1994; Daille, Gaussier & Lange, 1994; Daille, 1994; Daille, 1995; Frantzi & Ananiadou, 1996; Georgantopoulos & Piperidis, 1998).

Here, we are pursuing the issue of term DB organisation in a manner allowing for a more sound use and effective management of contained information.

Given the existing standards for organizing terminology resources, the main question raised in respect to the above, is how to create a DB or assimilated term list with properties supporting the efficient NLP treatment of texts, without falling within the restrictions posed by trivial word-list solutions. In this line, two more issues have to be considered:

- i) To determine adequate criteria for inclusion or not in the list of not yet fixed terms
- ii) To decide on what extend standardization principles may be kept by the above list in order to ensure its relevance both for practical applications and leadingedge technology development.

In respect to (i) above, a solution is to measure frequencies as against prominent vs. variant forms (allomorphs) of established terms, in order to decide on their inclusion in the list. A slight variation of this procedure is to count frequencies of appearance at a certain context and if rates reach or are very close to rates of already included terms, then the new ones are added to the list. It should be noticed that given the fact that our term DB is also intended to be used by human users, terms may be accompanied by various comments like: "older form", "occurs only in text group Y", etc. This allows us to add terms which are in the borderline between already established and *ad hoc* term status, or which are very likely to enter the established term list very soon without creating any problem for the MT system. This is achieved by marking these terms with a special flag which excludes the flagged items from been visible when the DB is used by the MT system, but it allows use by the human user.

As far as (ii) is concerned, it directly addresses the question "How will the term BD be structured?".

3. The structure of the term DB

Apart form theory-oriented approaches to standardisation, we used empirical observation in an effort to improve terminological annotation of the input text, so that the requirement for reusability of the DB and efficient performance by the systems making use of them, be best met. Having this specific goal, the design of our DB incorporated three main properties:

1) *The DB was thematically structured*. The terminology content of the input corpora was divided into thematic

domains/sectors so that the term DB was subdivided according to a thematic classification, which allowed the system to search in one or more thematic areas, by simply indicating the selected option before the actual search similar to the EURAMIS mechanism (Blatt, 1998).

- 2) *The DB was multi-use oriented.* Although, the here proposed DB structure was originally intended to improve an MT system performance, when linked with another application, the resulted DB can perform equally well. It can, for example, be used by human users as an advanced terminology lexicon, open to enrichment.
- 3) The DB functionality was based on a set of crosslinking indicators. Statistical methodologies for term extraction applied on the input texts, ended up with a list of candidate terms. The validated of the extracted terms as well as those terms which were not automatically recognised, were then entered and organized in BD-like tables providing cross-linked indications for the following properties:
- *Classification according to the head* of multi-word terms: This link allows all multi-word strings which share a common head to appear as a group in window if demanded by the user. For example, there is a link between all (multi-)word items contained in the DB for the head "expenditure" as shown next:

expenditure:

annual gross expenditures capital expenditure current expenditure gross domestic expenditure on R&D (GERD) gross national expenditure on R&D (GNERD)

•••

This function provides a grouping of (possibly) related terms both in respect to their syntactic and semantic content and facilitates term boundary definition either automatically or in the interactive mode.

- *Morpho-syntactic variants and synonyms*, especially for not yet fixed terms: In many cases there are more than one (1) forms noticed to indicate the same concept, as exhibited in the examples below. In the term DB all variants or synonyms for a term are mutually linked.
 - * man-year person-year
 - * patent disclosure disclosure of patent
 - s persons not in the labour force population not currently active
 - * support activity supporting activity
 - * unit surveyed reporting unit unit of observation (*statistic concept*)
- *Derivational morphology* properties of terms: In many cases, the termhood property of a stem is obtained by

a number of derivations. A specific link provides grouping of all terminology items, related via derivational word formation. The term list provided via this linking may contain also compounds which are related to a given stem. This function is exemplified by the following examples:

class-:

class classification classification system classify cross-classification cross-classify institutional classification subclass subclassification

industry :

civil industry computer services industry industry industrial industrial activity industrial application industrial innovation

• *Semantic-pragmatic relations* between concepts: This link connects otherwise unrelated items, which are grouped together because they form lists or closed sets inside a specific terminology domain. The following examples demonstrate the situation:

fields of study in S&T :

...

agricultural sciences engineering and technology humanities medical sciences natural sciences social sciences

economic indicators :

disputes earnings hours of work job

3.1. The data organisation

The parts of our terminology DBs intended to support MT are bilingual but not automatically reversible. This means that there are two different DBs (English-to-Greek and Greek-to-English) to be activated according to the selected source-target language pair.

Terms are registered accompanied by selected terminology information such as "domain of use/origin", "definition" and "example of use". The main organisation provides for two more fields: "variants" and "translations". Schematically, the available information can be represented as follows:

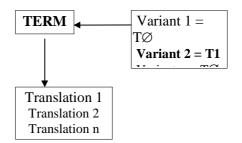


Figure 1 : term-variant / term-translation relations

The above allows for all variants of a term to be connected to its most outstanding form, but it implies that whenever further information about the specific term is required or translation takes place, this will be achieved by consulting the main term entry. A similar situation holds for the choice of the translation of the term. "Translation 1" is the most prominent form of the term in the target language and it is proposed (= solution automatically used) by the system unless the user selects one of the other translation alternatives. In this way, overgeneration is also controlled. We will illustrate by means of a concrete example. The English-Greek DB provides for the following links, where the source variant always makes reference to the prominent term form and translation automatically selects the respective target prominent term.

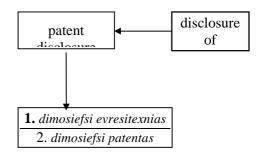


Figure 2 : English-to-Greek translation example

The screen for the same term in the Greek-English DB is as follows:

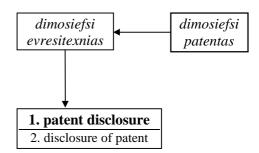


Figure 3 : Greek-to-English translation example

In Figure 1, we can see that the variants of a term may be related to possible translation options. The default case is that no preference for a translation other than the prominent choice (Translation 1) will be available. However, in the case where some translation is indicated in respect to a specific variant (illustrated in Figure 1, Variant 2), it will be automatically used by the system.

In order to extend use of the information available through the various cross-linking options underlying the entries of the DB, as indicated in the previous section, the user may activate the relevant links, offered as actionbuttons on his/her screen. The options available are either to view a panel with a list of items as is the output of the link associated with the derivational morphology properties of terms as well as the link of classification according to the head of multi-word terms, or to associate a term with some semantic/pragmatic property or morphosyntactic variant.

The whole nest of linkings is indicated in the following schema:

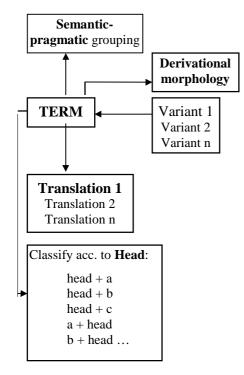


Figure 3 : Greek-to-English translation example

4. Conclusion

The organization of data which is proposed here, supports a more complete and efficient representation of complex relations and properties of terminology units and allows for linking options which result to reduction of unwanted overgeneration in MT because of reduction of ambiguity. The proposed linking nest improves functionality of the resources and enables better design or customization of term DBs according to user needs for monolingual as well as for multilingual applications. If a DB structured as described above, is combined with techniques for resolving problems of term recognition onprocessing, namely the syntax based multi word term recognition technique and the interactive term disambiguation indicated in the text, it makes part of an integrated system for terminology resolution, not only inside an MT context but also for use in a wide range of NLP applications.

5. References

- Blatt, A. (1998). "EURAMIS : Added value by integration", T&T Terminologie et Traduction, 1.1998, pp. 59-73.
- Bourigault, D. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In Proceedings of the 14th International Conference on Computational Linguistics.
- Church, K.W. & P. Hunks (1990). Word Association, Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1).
- Dagan, I. & K.W. Church (1994). Teright: Identifying and Translating Technical Terminology. In *Proceedings of* the EACL (1994).
- Daille, B., E. Gaussier & J.M. Lange (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING 94*, 515—521.
- Daille, B. (1994). Study and Implementation of combined techniques for automatic extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Languages*, Workshop at the 32nd Annual Meeting of ACL, Las Cruces, New Mexico.
- Daille, B. (1995). Combined approach for Terminology Extraction : Lexical statistics and linguistic filtering. TALANA, Université Paris 7.
- Efthimiou, E. (1992). Terminology and Machine Translation. In *Proceedings of Standardization of Terminology*, 218–228. TEE - ELOT.
- Efthimiou, E. & C. Alexandris (1998). On the treatment of extra-linguistic knowledge in grammar resources. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 1205–1210.
- Efthimiou, E. & M. Katsoyannou (2000). Handling terminology-intensive texts in Machine Translation. In *Proceedings of RIAO 2000* (in print).
- Frantzi, K.T. & S. Ananiadou (1996). Extracting nested collocations. In *Proceedings of COLING 96*, 41–46.
- Frantzi, K.T. & S. Ananiadou (1997). Automatic term recognition using contextual clues. In *Proceedings of Mulsaic* 97, IJCAI, Japan.
- Georgantopoulos, B. & S. Piperidis (1998). Eliciting Terminology knowledge for information extraction applications. In *Proceedings of EURISCO* 98.