Exploiting Technical Terminology for Knowledge Management

Fabio Rinaldi and Elia Yuste

Institute of Computational Linguistics,
University of Zurich, Switzerland,
email: {rinaldi, yuste}@ifi.unizh.ch

Abstract. In the world of globalization, it is essential for companies to be able to effectively manage their knowledge capital. Being capable to effectively create, store and retrieve institutional information is a crucial competitive advantage. Readily accessible Knowledge is needed in many business' aspect and tasks: support decision making, profile work processes, empower in-house knowledge workers (as well as external partners and clients). In this paper we focus on the importance of terminology management as one vital aspect within a corporate Knowledge Management strategy.

1 Introduction

Huge quantities of documents are created regularly in large organizations. Compared to ordinary Document Management systems, the focus of Content management systems (CMS) is upon individual content blocks, i.e. blocks of content are tagged with metadata and other attributes, which are held in a content database. By separating the management of content from its presentation (display), the task of maintenance and updating is significantly simplified. This is of particular relevance when a given content block is requested to appear on many corporate documents (passage leverage) or it is to be queried or manipulated by several users (user-defined knowledge variables) across the organization. More interestingly from our standpoint is the fact that a fully operational CMS will ideally interact with an array of corporate language resources (e.g. a term base) and applications (e.g. an answer extraction system, a summarization tool, etc.) within a business operational workflow. However, good CMS that look at corporate language resources are not yet commonplace, mainly because many organizations are not yet aware of the potential of their linguistic assets. This is precisely why we wish to concentrate on how to effectively access the knowledge residing in the organization's major language resource, a corpus of its specialized or technical documentation that has to be exploited, starting with meaningful terminology work.

We first discuss the pivotal role of terminology in technical domains and describe the operations adopted for structuring the terminology (section 2). Section 3 discusses how the gained structures can be exploited in the query process. Finally section 4 explores some related work.

2 Finding Structure in Terminology

The crucial importance of terminology in the process of Knowledge Management has long been recognized [1]. Any sort of technical documentation contains technical terminology that needs to be properly detected, managed and exploited before any NLP system can perform adequately. In our recent work we have considered two different types of technical texts: the Aircraft Maintenance Manual (AMM) of the Airbus A320 [2] and the GENIA corpus [3].

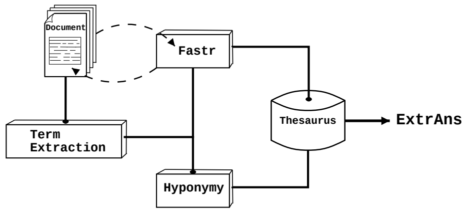


Fig. 1. Term Processing

The AMM, which in source form is approximately 120MB large, describes how the constituent parts of an Airbus A320 relate to each other, the testing and maintenance procedures for each part, as well as the tools and materials to be used. As 30% of the words in the running text belong to the terminology, pre-processing needs to be focused in this direction. Terminology extraction followed by thesaurus construction are necessary first steps before using the terms (figure 1).

We will not describe in detail in this paper the methodologies that were adopted for Terminology Extraction as they have been already presented in [4]. We will focus instead on the problem of discovering relations which are implicit in the extracted terminology, in particular synonymy, in order to conflate variants into a single synset, and hyponymy, with the aim of creating a taxonomy of synsets. The results of this phase of relation discovery is a taxonomy for the domain whose organizing element is the synset, each synset representing a domain specific concept. This process could be described as an attempt to elicit hidden knowledge, implicit in the domain.

Despite all efforts in standardization carried out by scientific boards, terminology standards associations and the like, it is often unavoidable that different technical writers use different (but related) surface forms to refer to the same domain concept. Besides, new technical developments will lead to the continuous creation of new terms. Even in consolidated sectors there are no absolutely reliable methods to enforce standardization. Consequently, when processing technical documents it is vital to recognize not only standardized terminology but also potential variations and possible new terms.

The process of terminological variation is well investigated [5, 6]. A subset of such variations identifies terms which are strictly synonymous. Our approach is based upon gathering these morpho-syntactic variations into units called synsets. The sets are defined by three weaker synonymy relations described in [7]. These

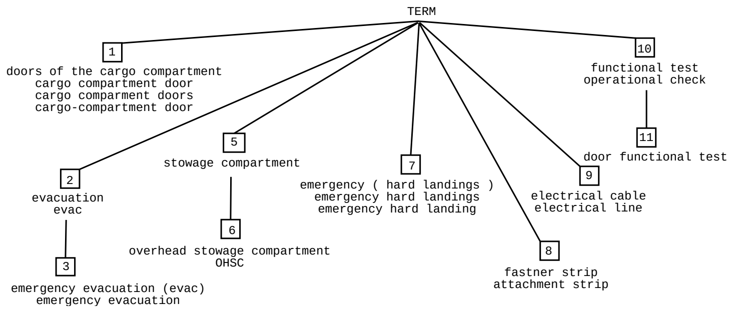


Fig. 2. A sample of the AMM computational thesaurus

synsets are then organized into a hyponymy (isa) hierarchy, a small example of which can be seen in figure (2).

The first stage is to normalize any terms that contain punctuation by creating a punctuation free version and recording that the two are strictly synonymous. Further processing is involved in terms containing brackets to determine if the bracketed token is an acronym or simply optional. In the former case an acronym-free term is created and the acronym is stored as a synonym of the remaining tokens which contain it as a regular expression. So **evac** is synonymous with **evacuation** but **ohsc** is synonymous with **overhead stowage compartment**. In cases such as **emergency (hard landings)** the bracketed tokens can not be interpreted as an acronym and so are not removed.

The synonymy relations are identified using the terminology tool Fastr [8]. All tokens of each term are associated with their part-of-speech[1], their morphological root[2] and their synonyms[3]. How tokens combine to form multi-token terms is represented as a phrasal rule, the token specific information carried in feature-value pairs. Metarules license the relation between two terms by constraining their phrase structures in conjunction with the morphological and semantic information on the individual tokens. We have designed the Metarules to identify strict synonymy that results from morpho-syntactic variation (**cargo compartment door ⟶ doors of the cargo compartment**), terms with synonymous heads (**electrical cable ⟶ electrical line**), terms with synonymous modifiers (**fastener strip ⟶ attachment strip**) and both (**functional test ⟶ operational check**). For a description of the frequency and range of types of variation present in the AMM see [4].

A simple algorithm determines lexical hyponymy between terms. Term `A` is a hyponym of term `B` if: `A` has more tokens than `B`, all the tokens of `B` are present in `A` and both terms have the same head. There are three provisions. First, ignore terms with dashes and brackets as **cargo compartment** is not

[1] as assigned by the IMS TreeTagger
[2] obtained from CELEX, http://www.kun.nl/celex
[3] as defined by WordNet, http://www.cogsci.princeton.edu/wn

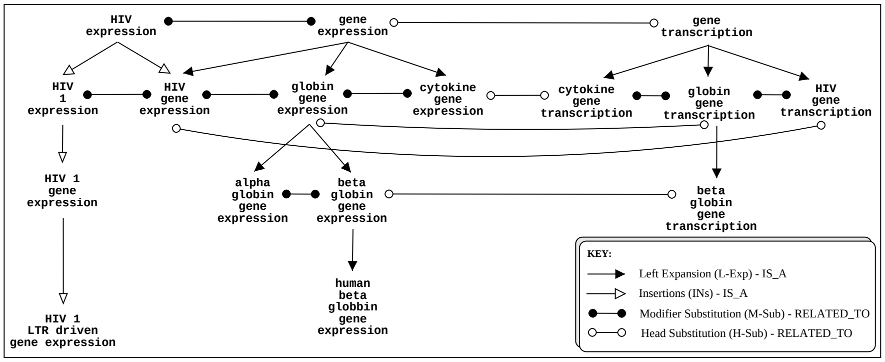

HIV expression
gene expression
gene transcription
HIV 1 expression
HIV gene expression
globin gene expression
cytokine gene expression
cytokine gene transcription
globin gene transcription
HIV gene transcription
HIV 1 gene expression
alpha globin gene expression
beta globin gene expression
beta globin gene transcription
human beta globbin gene expression
HIV 1 LTR driven gene expression
KEY:
Left Expansion (L-Exp) - IS_A
Insertions (INs) - IS_A
Modifier Substitution (M-Sub) - RELATED_TO
Head Substitution (H-Sub) - RELATED_TO

Fig. 3. Examples of Terminological Relations in the GENIA corpus.

a hyponym of **cargo - compartment** and this relation (synonymy) is already known from the normalization process. Second, compare lemmatized versions of the terms to capture that **stowage compartment** is a hyperonym of **overhead stowage compartments**. Finally, the head of a term is the rightmost non-symbol token (i.e. a word) which can be determined from the part-of-speech tags. This hyponymy relation is comparable to the insertion variations defined by [5]. Automatically discovering these thesaurus relations across 6032 terms from the AMM produces 2770 synsets with 1176 hyponymy links. Through manual inspection of 500 synsets 1.2% were determined to contain an inappropriate term. A similar examination of 500 hyponymy links verified them all as valid.

While the approach described so far exploits only 'endogenous' information (within the terms), different approaches have been proposed that make use of information explicitly provided by the author. We are currently experimenting with some of the patterns proposed by [9]. For example the following patterns can be used to identify NP_1 as an hyponym of NP_0:

such NP_0 as NP_1
NP_0 *including* NP_1

3 Using Terminology to enhance Information Access

The process of retrieving relevant information from corporate documentation necessarily begins with an user query. The query can be structured or unstructured, however it will most likely contain a reference to the domain concepts that the user is interested in. Such reference is typically expressed by means of a technical term. As it cannot be expected that the user of the system formulates a query using exactly the same wording that it is used in the background documentation, the system must be capable of detecting paraphrases and related terms, in order to locate relevant background documentation.

The problem resides in the imperfect knowledge of users of the systems, who cannot be expected to be completely familiar with the domain terminology. Even experienced users, who know very well the domain, might not remember the exact wording of a compound and use a paraphrase to refer to the underlying domain concept. Besides even in the manual itself, unless the editors have been forced to use some strict terminology control system, various paraphrases of the same compound will appear, and they need to be identified as co-referent.[4]

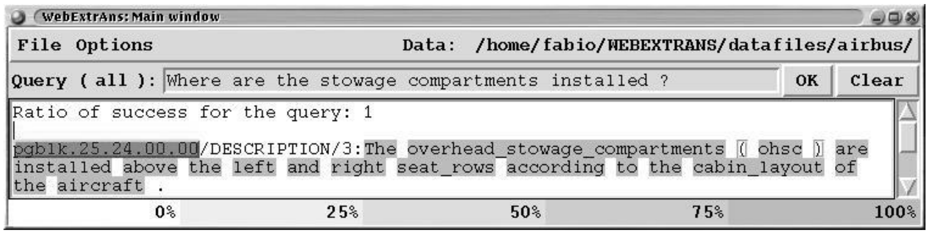
Discovering terminological relations, like synonymy and hyponymy, provides a crucial support in the process of query expansion (regardless of the specific type of query that is adopted by the system). As an example we describe the way terminology structure is exploited in our own Question Answering system (ExtrAns), specifically targeted at technical domains.

Processing is split into two distinct phases: the first offline step is Term Processing involving extraction and organization of the term thesaurus, as described in section 2. The next step, Linguistic Analysis, results in a semantic representation of the sentences – their **M**inimal **L**ogical **F**orm. These are stored along with their original location in a Knowledge Base. Online, the user query is processed using the same linguistic analysis, and the resulting MLF is matched against the Knowledge Base. The matches are then displayed in the document so users can contextualize these potential answers.

Part of the Linguistic Analysis involves the Link Grammar parser [11], generating a dependency structure for each syntactic interpretation of a single sentence. The multi-word terms from the thesaurus are identified and passed to the parser as single tokens. This prevents (futile) analysis of the internal structure of terms simplifying parsing by up to 50%.[5] This results in an average of 4.1 logical forms per sentence. Answers are identified by matching (logically proving) the query MLF against the MLFs stored in the Knowledge Base. During construction of the MLFs, thesaurus terms are replaced by their synset identifier. This results in an implicit 'terminological normalization' for the domain. The benefit to the QA process is an assurance that a query and answer need not involve exactly the same surface realization of a term. Utilizing the synsets in the semantic representation means that when the query includes a term, ExtrAns returns sentences that logically answer the query, involving any of the terms' synset members. When the thesaurus definition of terminological synonymy fails to locate an answer from the document collection, ExtrAns taps the thesaurus hyponymy relations. Instead of looking for synset members, the query is reformulated to include hyponyms and hyperonyms of the terms. For instance the query in figure 4 contains the domain term *stowage compartments* while the answer contains its hyponym term *overhead stowage compartments*.

[4] This problem has been described previously as the *paraphrase problem* [10].

[5] The measure refers to the average number of parses per sentence.



Data: /home/fabio/WEBEXTRANS/datafiles/airbus/

Query (all): Where are the stowage compartments installed ?

```
Ratio of success for the query: 1

pgblk.25.24.00.00/DESCRIPTION/3:The overhead_stowage_compartments ( ohsc ) are
installed above the left and right seat_rows according to the cabin_layout of
the aircraft .
```

Fig. 4. overhead_stowage_compartment is an hyponym of stowage_compartment

4 Related Work

Within the medical domain, the Unified Medical Language System (UMLS), created by the National Library of Medicine[6] collects terminologies from differing sub-domains in a metathesaurus of concepts. The organization of the terms involve hyponymy and lexical synonymy, which is the same approach that we have followed in our activity (though on a much smaller scale). An application of the UMLS resource is PubMed[7] which retrieves the abstracts from medical journals by relating metathesaurus concepts against a controlled vocabulary used to index the abstracts.

Many Information Extraction (IE) tasks over this domain utilize the UMLS terminology in conjunction with shallow parsing in the construction of knowledge bases. A statistical *bag-of-words* approach applied at the sentence level [12] determines predicate relations between proteins and chemicals, as long as multi-word terms are identified in the *bag*. Syntactically identifying object-predicate-object relations [13] would be impossible without the prior identification of multi-word term objects in the Metathesaurus. Inferences have also been directly extracted from the occurrence of terminology under certain of the MeSH headings [14]. A term X under the abstract heading *methods*, and term Y under *diagnosis* implies that X *diagnoses* Y.

Hamon & Nazarenko [7] explores the terminological needs of consulting systems. This type of IR guides the user in query/keyword expansion or proposes various levels of access into the document base on the original query. A method of generating three types of synonymy relations is investigated using general language and domain specific dictionaries.

5 Conclusion

In this paper we have focused on the importance of identifying and structuring domain terminology in the context of Intelligent Information Access. Tools that effectively exploit domain terminology for providing access to technical documentation are a key factor of advanced Knowledge Management applications.

[6] http://www.nlm.nih.gov/research/umls/

[7] http://www.ncbi.nlm.nih.gov/pubmed

References

1. Mayer, F., Schmitz, K.D., Zeumer, J., eds.: Terminologie und Wissensmanagement. (2004) Köln, 26-27 März.
2. Rinaldi, F., Hess, M., Dowdall, J., Mollá, D., Schwitter, R.: Question answering in terminology-rich technical domains. In Maybury, M., ed.: New Directions in Question Answering. MIT/AAAI Press (2004) .
3. Kim, J., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics **19** (2003) 180–182
4. Rinaldi, F., Dowdall, J., Hess, M., Kaljurand, K., Koit, M., Vider, K., Kahusk, N.: Terminology as Knowledge in Answer Extraction. In: Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02), Nancy (2002) 107–113 .
5. Daille, B., Habert, B., Jacquemin, C., Royauté, J.: Empirical observation of term variations and principles for their description. Terminology **3** (1996) 197–258
6. Ibekwe-Sanjuan, F.: Terminological Variation, a Means of Identifying Research Topics from Texts. In: Proceedings of COLING-ACL, Quebec,Canada (1998) 571–577
7. Hamon, T., Nazarenko, A.: Detection of synonymy links between terms: Experiment and results. In Bourigault, D., Jacquemin, C., L'Homme, M.C., eds.: Recent Advances in Computational Terminology. John Benjamins Publishing Company (2001) 185–208
8. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. MIT Press (2001)
9. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING '92, Nantes. (1992) 539–545
10. Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., Mollá, D.: Exploiting paraphrases in a question answering system. In: The ACL-2003 workshop on Paraphrasing (IWP2003), July 2003, Sapporo, Japan. (2003) .
11. Sleator, D.D., Temperley, D.: Parsing English with a link grammar. In: Proc. Third International Workshop on Parsing Technologies. (1993) 277–292
12. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text source. In: the Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg,Germany. (1999) 77–86
13. Sekimizu, T., Park, H., Tsujii, J.: Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. Genome Informatics, Universal Academy Press. (1998)
14. Cimino, J., Barnet, G.: Automatic Knowledge Acquisition from Medline. Methods of Information in Medicine **32** (1993) 120–130