Exogeneous and Endogeneous Approaches to Semantic Categorization of Unknown Technical Terms

Farid Cerbah
Dassault Aviation - DPR/DESA - 78, quai Marcel Dassault
92552 Saint-Cloud cedex 300 - France
farid.cerbah@dassault-aviation.fr

Abstract

Acquiring and updating terminological resources are difficult and tedious tasks, especially when semantic information should be provided. This paper deals with *Term Semantic Categorization*. The goal of this process is to assign semantic categories to unknown technical terms. We propose two approaches to the problem that rely on different knowledge sources. The exogeneous approach exploits contextual information extracted from corpora. The endogeneous approach relies on a lexical analysis of the technical terms. After describing the two implemented methods, we present the experiments that we conducted on significant test sets. The results demonstrate that term categorization can provide a reliable help in the terminology acquisition processes.

1 Introduction

Terminological resources have been found useful in many NLP applications, including Computer-Aided Translation and Information Retrieval. However, to have a significant impact on applications, terminological knowledge should not be limited to flat nomenclatures of single-word and multi-word terms. They should include semantic knowledge, such as semantic categories and various kinds of semantic relations (hyperonymy/hyponymy, synonymy). This paper focuses on the assignment of semantic categories to technical terms. Semantic categories may play a crucial role in many applications, and particularly when disambiguation processes are required. For example, in our applicative framework, semantic categories ensure coarse sense division of polysemous terms and are actually used to improve the French-to-English translation of the technical documentation. Assigning semantic categories to technical terms is a difficult and time-consuming task. Highly specialized skills are required and, even though the major concepts represented in these terminological resources pertain to the aeronautic domain, various related knowledge areas are concerned, such as *Data Processing*, *Mechanics*, and *Manufacturing Processes*. Our goal is to elaborate a tool that helps terminologists to assign semantic categories when updating the reference terminology. We think that significant support can be provided to the terminologists by taking advantage of existing categorized terms and their usages in documents. Such a tool can be integrated within a terminology acquisition environment as a complement to a term extraction component.

We distinguish two kinds of approaches to semantic categorization. In a way similar to corpus-based methods for Word Sense Disambiguation (WSD)(Yarowski, 1992; Ide and Véronis, 1998), an *exogeneous approach* exploits contextual information extracted from corpora in order to determine the most plausible categories. By contrast, an *endogeneous approach* relies solely on a lexical analysis of multi-word terms which are very frequent in terminological databases. This approach is based on the assumption that lexical units used in the composition of technical terms are relevant indicators of semantic domains.

The rest of the paper is organized as follows. Section 2 describes the terminological resources used in this study. Section 3 compares term categorization with related issues, such as thesaurus extension, WSD and term clustering. Sections 4 and 5 describe the two proposed methods. Results and evaluation are given and discussed in section 6. Directions for further research are pointed out in last section.

French	**English**	**Categories**	**POS**
anti-dérapage à long terme	long-action antislip	NAV	N
assiette de consigne au décollage	required takeoff attitude	CKI	N
barre de remorquage	tow bar	TOO	N
dérive	fin	STR	N
dérive	wander	PRO, FLP	N
dériver	to unrivet	MEC,MAI	V
dérivée	derivative	COM	N
embout coulissant	sliding endpiece	ENG,LGR	N
enregistreur de fatigue	fatiguemeter	FPL	N
enregistreur de paramètres	flight data recorder	FPL	N
jeu de protecteurs boudin cabine	set of cockpit seal protectors	TOO	N
bit d'appui touche	keystroke bit	DPR	N
amplificateur téléphone de bord	flight crew interphone amplifier	RTL	N
ne pas déboucher	to be blind	MAI	LV

- **MAI**: Maintenance, **NAV**: Navigation, **CKI**: Cockpit Indications, **TOO**: Tools, **FLP**: Flight Parameters, **ENG**: Engines, **LGR**: Landing Gear, **STR**: Aircraft Structure, **DPR**: Data Processing, **RTL**: Radiocommunications.

Table 1: A sample of the terminological database.

2 The Terminological Resources

We use in this study a French/English bilingual terminology of the aeronautic domain. This hand-crafted database results from a multi-disciplinary effort involving technical writers, translators, terminologists and engineers. In its current state, the database contains 12,267 French/English term couples, structured in 70 semantic categories. As already observed in several terminological databases, multi-word terms cover the larger part of the database (nearly 80%). The described terms are mostly nouns but the database also contains about 200 verbs and verbal phrases. Table 1 gives some examples of terms and a short description of the associated categories. These linguistic resources are integrated in a computer-aided translation environment used by technical writers.

Semantic categories have originally been introduced in order to distinguish the various senses of polysemous terms. Each term couple is annotated with one or more categories specifying the contexts in which the translation is recommended. An entry is associated to a term for each identified meaning. For example, the french term *dépassement* has at least two possible meanings, corresponding to two different translations: *overflow* in the *Data Processing* category (DPR) and *out-of-flushness* in the *Aircraft Structure* category (STR). As shown in the examples of table 1, the assignment of semantic categories has been extended to monosemous terms.

In our experiments of term categorization, only the french terms have been used.

3 Related Work

Term Semantic Categorization is on several aspects similar to *Thesaurus Extension* (Uramoto, 1996; Tokunaga et al., 1997). Our methods are close to those used for positioning unknown words in thesauri. However, the two issues can be differentiated with respect to the manipulated data. A thesaurus is intended to cover a large set of conceptual domains while a terminological database is focused accurately on a specific topic and its related domains. For example, in (Tokunaga et al., 1997), the thesaurus to be extended contains more than 500 categories. This tends to make the problem harder, but, since many categories are strictly independent, it is easier to find distinctive features between categories. By contrast, our terminological database contains only 70 categories. But, in this restricted set, we find categories corresponding to close or even overlapping knowledge areas. It is more difficult to differentiate them.

Furthermore, the endogeneous approach, which exploits the multi-word nature of terminological units, cannot be applied to thesaurus extension because of the large amount of single-word thesaurus entries[1].

It is useful to compare exogeneous term categorization with corpus-based WSD methods. In both cases, contextual information extracted from corpora are used in order to assign the most plausible semantic tags to words. In WSD, the contextual cues that co-occur with the *target word* constitute the main training source whereas, in term categorization, the contextual information occurring with the term to be categorized should not be included in training data since the term is supposed to be unknown. The only relevant training sources are the contextual cues surrounding the already categorized terms. This is a basic difference that explains why WSD tasks usually achieve better performance than term categorization and thesaurus extension.

In a terminology acquisition framework, Habert et al. (1998) propose an exogeneous categorization method of unknown simple words. They use local context of simple words provided by a term extraction system.

Endogeneous term categorization can also be compared with some approaches to term clustering (Bourigault and Jacquemin, 1999; Assadi, 1997). These approaches take advantage of the lexical and syntactic structures of technical terms in order to build semantic clusters.

4 Exogeneous Categorization

We tested several classification models. Our first experiments were carried out with Example-based classifiers. We used our own implementation of K-nearest neighbors algorithm (kNN), and then the *TiMBL* learner (Daelemans et al., 1999), which provides several extensions to kNN, well-suited for NLP problems. Nevertheless, in the current state of our work, better results were obtained with a probabilistic classifier similar to the one used by (Tokunaga et al., 1997) for thesaurus extension. Due to lack of space, only this method will be described in this paper.

We use as contextual cues the open-class words (nouns, verbs, adjectives, adverbs) that co-occur in the corpus with the technical terms. More precisely, the cues are open-class words surrounding the occurrences of the term in some window of predefined size. Each new term to be categorized is represented by the overall set of contextual cues that have been extracted from a part of the corpus (test corpus).

4.1 Probability Model

Let us consider a term T for which the contextual cues $\{w_i\}_{i=1}^{n}$ have been collected in the test corpus. The categorization of this term amounts to find the category C^* that maximizes probability $P(C|T)$:

$$C^* = \arg\max_C P(C|T) \qquad (1)$$

According to the exogeneous approach, the probability that a term T belongs to category C depends on the contextual cues of T:

$$C^* = \arg\max_C \sum_{i=1}^{n} P(C|w_i)P(w_i|T) \qquad (2)$$

After applying Bayes'rule:

$$C^* = \arg\max_C P(C) \sum_{i=1}^{n} \frac{P(w_i|C)P(w_i|T)}{P(w_i)} \qquad (3)$$

The probabilities of the equation 3 are estimated from training data:

- $P(w_i|C)$ is the probability that a word w_i co-occurs with a term belonging to category C. It is estimated in the following way:

  $$P(w_i|C) = \frac{N_w(w_i, C)}{\sum_{w_j} N_w(w_j, C)} \qquad (4)$$

  $N_w(w_i, C)$ is the number of times that w_i co-occurs with a term belonging to category C.

  This probability accounts for the weight of cue w_i in category C.

[1] For Japanese, (Tokunaga et al., 1997) reports some promising experiments of endogeneous categorization to thesaurus extension. The approach relies on properties of Japanese word formation rules and, thus, it can hardly be adapted for other languages. Their experiments suggest that exogeneous and endogeneous approaches are complementary.

- $P(w_i|T)$ is the probability that w_i co-occurs with T:

$$P(w_i|T) = \frac{N_w(w_i,T)}{\sum_{w_j} N(w_j,T)} \qquad (5)$$

  $N_w(w_i,T)$ is the number of times that w_i co-occurs with T.

- $P(w_i)$ is the probability of cue ${w_i}^2$:

$$P(w_i) = \frac{N_w(w_i)}{\sum_{w_j} N_w(w_j)} \qquad (6)$$

- P(C) is the prior probability that a term of the corpus belongs to the category C:

$$P(C) = \frac{N_t(C)}{\sum_{C_i} N_t(C_i)} \qquad (7)$$

  where $N_t(C)$ is the occurrence number in training data of terms belonging to C. This probability accounts for the weight of category C in the corpus.

4.2 Training and Test

The exogeneous classifier starts with the selection of *test documents* in the corpus. Technical terms found in these documents will form the *test set*. The remaining documents represent the training corpus. Training and test stages are the following:

- **POS tagging.** The test and training corpora are tagged with MultAna, a tagger designed as an extension of the Multex morphological analyzer (Petitpierre and Russell, 1995). Occurrences of the technical terms are identified during this stage and the terms to be categorized are those which are identified in the test corpus.

- **Extraction of contextual cues.** For each term occurrence in training and test data, the contextual cues are collected. Only the lemmas of open-class words are used and cues may correspond to multi-word terms. Each test term is then represented by the set of cues which have been collected in test data.

Incrémentation (*Increment*) **[DPR]**
DPR (0.2106), CKI (0.2027), RDR (0.1967)

Décrémentation (*Decrement*) **[DPR]**
DPR (0.2111), CKI (0.1843), RDR (0.1654)

Renseigner (*Inform*) **[NAV]**
CKI (0.24), VOR (0.21), DPR (0.1276)

Entrée d'air (*air intake*) **[ENG]**
ENG (0.1895), FUE (0.1214), DOC (0.1192)

Moteur (*Engine*) **[ENG]**
ENG (0.1494), CDV (0.1285), DOC (0.1095)

Effacement de données (*Data clearing*) **[RTL]**
DPR (0.2059), DOC (0.1357), RTL (0.129)

Téléphone de piste (*Ground telephone*) **[RTL]**
RTL (0.1251), ELE (0.1131), EQX (0.1011)

Figure 1: Some results of the exogeneous categorization.

- **Frequency calculation and probability estimation.** Training data are explored to compute the frequencies (occurrences and co-occurrences) of cues, terms and categories. As mentioned earlier (section 3), the cue occurrences which have been collected around the test terms are ignored during this step. The probabilities required for the categorization operation are then computed.

- **Categorization of the test terms.** The most probable categories are assigned to each test term (see section 4.1). Figure 1 gives some examples of exogeneous categorization[3].

5 Endogeneous Categorization

Our approach to endogeneous categorization is simpler. It is exclusively based on a quantitative analysis of the lexical composition of technical terms. Henceforth, the open-class words used to compose technical terms will be called *terminological components*. The endogeneous approach relies on a much more restricted source of data than the exogeneous approach, since the component set of a terminological database is quantitatively limited compared with the set of contextual cues extracted from corpora. Nevertheless, we make the assumption that this quantitative limitation is partly compensated by the

[2] Note that the categorization process could be simplified by eliminating $P(w_i)$, since this quantity is constant for all categories.

[3] Some category labels are described in table 1.

ARR		MAP	
architecture (*architecture*)	8.440	fonderie (*casting*)	7.910
encadrement (*framing*)	8.440	placage (*cladding*)	7.872
mousse (*foam*)	7.331	schoopage (*schoop process*)	7.232
capotage (*fairing*)	7.445	drapage (*layup*)	7.111
châssis (*rack*)	7.353	cuisson (*baking*)	6.912
élastique (*elastic*)	7.020	moule (*mould*)	6.838
cloison (*bulkhead*)	6.632	compacter (*to compact*)	6.591
escamotable (*retractable*)	6.441	broche (*pin*)	6.591
suspension (*shock mount*)	6.266	chimique (*chemical*)	6.134
ventilation (*ventilation*)	6.114	usinage (*milling*)	6.101

Table 2: The top ten most significant terminological components of categories ARR (*Arrangement*) and MAP (*Manufacturing Processes*).

strong discrimination power of terminological components.

The training phase assigns to each category a set of representative components with respect to some association score. The categorization phase determines the most plausible categories of a term according to its components.

5.1 Association Score

To estimate the dependency between components and categories, we experimented several association criteria. The choice of these criteria has been influenced by the comparative study described in (Yang and Perdersen, 1997) on feature selection criteria for text categorization. We tested several measures, including component frequency, information gain and mutual information. Our best results were achieved with mutual information which is estimated using:

$$I(w, C) \approx \log_2 \frac{N_w(w, C) \times N_w}{N_w(C) \times N_w(w)} \qquad (8)$$

- $N_w(w, C)$ is the frequency of component w in category C.
- N_w is the total number of component occurrences.
- $N_w(C)$ is the total number of component occurrences in category C. This factor reduces the effect of the components weakly represented in category C, compared with the other components of C.
- $N_w(w)$ is the frequency of component w in the terminological database. This factor reduces the effect of the components that denote basic concepts spread all over the database. For example, the components *speed, altitude, pressure* have high frequencies in category FLP (*Flight Parameters*), but, as basic concepts, they also appear frequently in many categories.

Table 2 gives for two categories the ten most representative components according to this score.

The association score between a term T (with components $\{w_i\}_{i=1}^{n}$) and a category C is given according to the components of T:

$$A_t(T, C) = \frac{N_t(C)}{N_t} \sum_{i=1}^{n} I(w_i, C) \qquad (9)$$

$N_t(C)$ is the number of terms pertaining to category C and N_t is the total number of terms. The factor $\frac{N_t(C)}{N_t}$ favors larger categories.

The categorization task determines the category C^* that maximizes the association score:

$$C^* = \arg\max_{C} A_t(T, C) \qquad (10)$$

5.2 Training and Test

Only multi-word terms can be categorized with this method since our endogeneous approach is by nature not relevant for simple words. A test set of compound terms is extracted from the terminological database. The remaining terms are used for training. The training terms are analyzed in order to assign to each category its terminological components. Then, component frequencies and association scores are computed.

	#T	k=1	k=2	k=3	k=4	k=5
	312	45.07	68.22	79.15	88.10	91.40
	89	42.25	73.14	86.95	91.57	98.19
	98	50.65	84.94	91.11	94.93	96.96
	120	48.11	78.95	91.26	96.83	97.92
	253	59.17	81.98	92.35	96.75	98.65
	125	73.16	89.30	96.85	98.68	99.59
	234	68.19	87.63	94.14	96.99	98.31
	205	49.05	83.86	95.12	98.61	99.47
	203	47.25	72.18	85.66	94.40	98.48
	105	44.55	69.61	87.28	93.97	97.11
Tot.	1744					
Avg.		52.75	78.98	89.99	95.08	97.61

Table 3: Results for the exogeneous approach. Ten experiments have been run for a total test set of 1744 terms.

	#T	k=1	k=2	k=3	k=4	k=5
	229	47.96	62.05	71.23	75.77	78.50
	232	42.60	56.38	62.28	69.73	74.23
	228	41.5	57.89	63.03	72.76	76.27
	227	45.54	61.05	68.57	76.74	80.18
	239	46.38	61.01	67.93	73.41	75.94
	229	43.41	57.54	65.42	71.16	75.34
	231	43.75	60.74	70.04	73.97	77.62
	228	42.85	60.18	66.97	70.50	73.85
	237	45.68	58.62	66.37	72.10	74.67
	240	42.30	58.82	68.48	70.29	74.05
Tot.	2320					
Avg.		44.19	57.77	67.03	72.64	76.06

Table 4: Results for the endogeneous approach. Ten experiments have been run for a total test set of 2320 terms.

During the test phase, each test term is annotated with the most plausible categories according to its components.

6 Experiments and Evaluation

To estimate the accuracy of the exogeneous method, we used a domain-specific corpus of 541,964 words, composed of documents pertaining to various textual genres (software specifications, maintenance procedures, manufacturing notices...). This corpus covered 63 of the 70 categories. Each run starts with the selection of a document among the corpus documents. The known terms identified in this document are considered as test terms. We used relatively wide contexts. The cues were extracted in a window of ±20 words around the term. Each run involved more than 70,000 contexts of term occurrences. To experiment the endogeneous approach, test sets of compound terms have been randomly extracted from the terminological database.

We adopted an evaluation scheme similar to that defined in (Tokunaga et al., 1997) for thesaurus extension. The categorization is considered successful if the right category appears among the k first categories assigned by the classifier. Within a semi-automatic acquisition framework, this evaluation scheme is more suitable than strict evaluation where only the first category assigned by the classifier is considered as relevant (evaluation restricted to k=1)[4]. From our application perspective, it is useful to provide to the terminologist a restricted set of less than 5 plausible categories instead of the complete set of 70 categories without prior filtering.

In the experiments described in (Tokunaga et al., 1997), k takes the values 5, 10, 20 and 30 for averaged performance ranging from 26.4% to 55.9% (the choice is made among 544 categories). Some of their precise experiments yielded an accuracy greater than 80% for k = 30. In our experiments, we measured accuracy for k=1 to 5. Some results are given in tables 3 and 4. The scores are higher than those achieved in thesaurus extension, especially with the exogeneous approach (from 52.75% to 97.61%). We should however keep in mind that we deal with a different kind of data (see section 3).

7 Conclusion and Further Work

We have presented in this paper two approaches to term semantic categorization that have been fully implemented and experimented on significant test sets. The results achieved in this work demonstrate that term categorization tasks could be integrated within a semi-automatic

[4] Some limitations of strict evaluation are also pointed out in (Resnik and Yarowski, 1999).

terminology acquisition process to provide an active support to terminologists.

The solutions to this problem can be considerably improved and we have identified several promising directions for further research.

Our experiments show that exogeneous categorization is noticeably the most efficient of both approaches. However, it requires much more knowledge sources and computational overhead. It is more exposed to data sparseness, since large amounts of contextual data are not always available, especially in technical domains. We should stress that this study benefited from the availability of a highly relevant corpus. This means that, for sake of robustness, other methods (even less efficient) and relevant knowledge sources should not be neglected. The two proposed approaches are complementary in the sense that they take advantage of distinct knowledge sources. Further work will investigate the various ways to combine them in order to improve the overall performance.

The use of relational information, and particularly syntactic relations, is another major direction for further research. Exogeneous categorization is based on a *bag of words/lemmas* model since wide contexts of lemmatized words were used, without consideration for the positions of these cues and their potential syntactic relationships with the target terms. Syntactic information extracted from local context, as verb-object relations, is another major source for exogeneous categorization that has been exploited in thesaurus extension methods. The endogeneous approach can also be improved by exploiting the syntactic structure of the technical terms. In our approach, all components of technical terms are equally weighted, independently of their syntactic roles within the terms. More accurate association scores can be introduced by taking advantage of head/modifier relations.

Finally, we should note that the bilingual nature of our terminological resources has not been taken into account. Minor changes are required to make the two classifiers work for English. Further experiments will be conducted on the English resources. In this bilingual context, either the French or the English expression (or both) could be used to categorize a given term.

References

H. Assadi. 1997. Knowledge Acquisition from Texts : Using an Automatic Clustering Method Based on Noun-Modifier Relationships. In *35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

D. Bourigault and C. Jacquemin. 1999. Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99)*, pages 15–22, Bergen.

W. Daelemans, J. Zavrel, K. van del Sloot, and Antal van den Bosch, 1999. *TiMBL: Tilburg Memory Based Learner – Version 2.0 – Reference Guide.* Tilburg University.

B. Habert, A. Nazarenko, P. Zweigenbaum, and J. Bouaud. 1998. Extending an Existing Specialized Semantic Lexicon. In *Proceedings of first International Conference on Language Resources and Evaluation*, pages 663–668, Granada.

N. Ide and J. Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40.

D. Petitpierre and G. Russell. 1995. MMORPH – The Multext Morphology Program. Technical report, Multext Deliverable 2.3.1.

P. Resnik and D. Yarowski. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sens Disambiguation. *Natural Language Engineering*, 5(2):113–133.

T. Tokunaga, A. Fujii, and M. Iwayama. 1997. Extending a thesaurus by classifying words. In Association for Computational Linguistics, editor, *ACL Workshop on Automatic Information Extraction and Building of Lexical semantic Resources*, pages 16–21.

N. Uramoto. 1996. Positioning unknown words in a thesaurus by using information extracted from a corpus. In *Proceedings of Coling 96*, pages 956–961.

Y. Yang and J. Perdersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*,.

D. Yarowski. 1992. Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of Coling 92*, Nantes.