Disambiguating automatic semantic annotation based on a thesaurus structure

Véronique Malaisé[1] Luit Gazendam[2] Hennie Brugman[3]
(1) Vrije Universiteit, Amsterdam
(2) Telematica Institute, Enschedé, Netherlands
(3) Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
vmalaise@few.vu.nl, Luit.Gazendam@telin.nl, Hennie.Brugman@mpi.nl

Résumé La relation *voir/employé pour* d'un thesaurus est souvent plus complexe que la (para-)synonymie recommandée dans le standard ISO-2788 qui décrit le contenu de ces vocabulaires contrôlés. Le fait qu'un non descripteur puisse être rattaché à plusieurs descripteurs, et le fait que seuls ces derniers soient pertinents dans le cadre de l'indexation contrôlée, font que cette relation est complexe à utiliser dans un contexte d'annotation automatisée : elle génère des cas d'ambiguité. Dans ce papier, nous présentons CARROT, un algorithme que nous avons mis au point pour classer les résultats de notre chaîne de traitements pour l'Extraction d'Information, et son utilisation dans le cadre de la sélection du descripteur correct parmi plusieurs possibilités. Cette sélection est fait dans le but d'une validation humaine, afin de simplifier et d'accélérer le travail des documentalistes annotant au quotidien et se base sur la structure de leur thesaurus. Nous arrivons à un succès de 95.5 %, et discutons ces résultats ainsi que des perspectives à cette expérimentation.

Abstract The *use/use for* relationship a thesaurus is usually more complex then the (para-)synonymy recommended in the ISO-2788 standard describing these controlled vocabularies content. The fact that a non preferred term can refer to multiple preferred terms, and the fact that only the latter is relevant in controlled indexing, makes this relationship difficult to use in automatic annotation applications : it generates ambiguity cases. In this paper, we present the CARROT algorithm, meant to rank the output of our Information Extraction pipeline, and how this algorithm can be used to select the relevant preferred term out of different possibilities. This selection is meant to provide suggestions of keywords to human annotators, in order to ease and speed up their daily process and is based on the structure of their thesaurus. We achieve a 95.5 % success, and discuss these results along with perspectives for this experiment.

Mots-clefs : Desambiguisation sémantique, algorithme de classement, Annotation automatisée

Keywords: Word Sense Disambiguation, Ranking algorithm, Automatic annotation

1 Introduction

Thesauri are controlled vocabularies, often used for indexing and retrieving documents from collections. The standard thesauri contain two types of elements, preferred and non preferred terms, related with a link called *use/use for*. This link is considered as (para-)synonymy in the ISO-2788 standard (ISO, 1986) and can thus be useful for (semi-) automatic indexing applications : it can enable to index a document with a preferred term (which is the type of thesaurus based controlled annotation we are interested in) either if the document contains an occurrence of the preferred term or if it contains occurrences of the corresponding non preferred term. In reality, this *use/use for* relationship is often more complex, and can generate ambiguity problems when used "as is" in an automatic application. We present in this paper the solution that we have developed in our project for selecting the relevant preferred term, given an occurrence of an ambiguous non preferred term in a text. This selection algorithm is based on the thesaurus's structure. The thesaurus we used in this experiment is the GTAA, which is employed for indexing and retrieving TV programs at the Netherlands Institute for Sound and Vision, the Dutch national TV archives. Our project, CHOICE[1], is collaborating with this Institute and focuses on easing and speeding up the work of cataloguers by providing them with a ranked set of keywords referring to their thesaurus' entries as indexing suggestions. We will present our project's goal and the specificity of this use case in the following section (section 2), followed by a description of thesauri in general and the GTAA itself (section 3). In this section, we will show the different semantics of the *use/use for* relationships and the problem of having multiple links between preferred and non preferred terms. We then present our annotation pipeline (section 3.4), including the algorithm that we elaborated to rank the keywords extracted by our pipeline, and that we propose here for selecting the relevant preferred term given an occurrence of an ambiguous non preferred term in a text (section 3.5). Section 5 shows our experiment to evaluate this algorithm on this disambiguation problem. We achieved a 95.5 % of success, but are still facing minor and more important problems. We discuss them and conclude with perspectives for this experiment in section 6.

2 The CHOICE project

Charting the Information Landscape Employing Context Information, the CHOICE project deals with the suggestion of metadata from textual resources to annotate documents. In the context of the Dutch TV archives, the cataloguers check a set of textual documents, on top of watching the program itself, to make their description. One of the goals of our project is to build on existing Information Extraction platforms, extend and tune them to our specific needs in order to cope with the particularities of this specific use case and provide the cataloguers with a relevant set of keywords as indexing suggestions. Our Information Extraction is based on the content of the thesaurus that they are currently using at Sound and Vision, enriched and transformed by us. We present this thesaurus in the following section, and the specificity of our task in the section describing our ranking algorithm.

[1] `http://www.nwo.nl/CATCH/CHOICE`

3 The GTAA thesaurus

3.1 A thesaurus according to the ISO 2788 standard

A thesaurus is *The vocabulary of a controlled indexing language, formally organized so that the* a priori *relationships between concepts (for example as "broader" and "narrower") are made explicit.*[2]

Although this definition mentions *concepts*, a thesaurus contains terms (preferred and non preferred terms), organized according to 5 relationships : broader term (BT), narrower term (NT), related term (RT), use (US[3]) and use for (UF). A preferred term is *A term used consistently when indexing to represent a given concept* [...], whereas a non preferred term should not be used for indexing, but is only useful at search time to point different words possibly expressing the same idea to the one that has been chosen to represent it in the thesaurus. The other relationships should stand only between preferred terms. BT relates a term with one more generic then itself, supposed to index a larger set of documents. For example *Means of transportation* is a BT of *Bus*. NT is the relationship between a term and a more specific one, that should be used to index a subset of the documents indexed by the more generic one (*Bus* and *School bus*). RT is a non hierarchical relationship between two terms in the same domain, as *Bus* and *Driver*, for example.

3.2 The GTAA

The GTAA thesaurus, a Dutch acronym for "Common Thesaurus for Audiovisual Archives", is the primary source for vocabulary used in the Sound and Vision documentation process. It contains approximately 160.000 terms. They are divided in 6 disjoint facets : Keywords (about 3800 preferred terms), Locations (about 14.000), Person Names (about 97.000), Organization-Group-Band Names (about 27.000), Maker Names (about 18.000) and Genres (113 preferred terms). The thesaurus mainly uses constructs as presented in the ISO 2788 standard and commonly used in companies or institutions : amongst others, use, use for, broader term, narrower term, related term. Terms from all facets of the GTAA may have related terms and use for relationships, but only Keywords and Genres can also have broader term/narrower term relations, organizing them into a set of hierarchies. Additionally, Keyword terms are thematically classified in 88 subcategories of 16 top Categories (Nature, Society,...). Although the data model that is used for the thesaurus allows links between terms across facets, no instances of these links currently exist. This experiment concerns only automatic indexing with terms from the Keyword facet.

3.3 Different semantics and non uniqueness of the Use relationship

In the GTAA, there are 1377 US relationships, *i.e.* 1377 times a non preferred term is associated with a preferred term. Some of these non preferred terms are associated with multiple different preferred terms and some of the preferred term are associated with multiple non preferred terms. In the first case, the non preferred terms is polysemic or can be considered in two different

[2] Quotation from (ISO, 1986), section 3-Definitions.

[3] In general this relationship is encoded USE, but this acronym is the one used in the GTAA.

domains, each meaning or domain having an explicit preferred term : for example, the non-preferred term minority[4] has two preferred terms, ethnical minority and religious minority. In the latter case (one preferred term associated with different non preferred terms), different notions were grouped under one common and single preferred term, for sake of easing the thesaurus' use (the fewer terms there are, the easiest it is to find the most appropriate one when indexing) or because the distinction was not relevant for indexing the TV programs : for example, the preferred term diplomats groups two non preferred terms, ambassadors and consuls). As a matter of fact, when we look into the nature of the US, UF relationship we see four different types :

- Synonyms : To cleanse US To clean
- Meronym : Sabbath US Jewish religion
- Hyponym : Scanner US Hardware
- Semantically related : Geiger counter US radioactivity

83 non preferred terms are associated with more then one preferred term in the thesaurus, ranging from 2 to 3 different preferred terms. This non unique association can be a source of problems when using the thesaurus' content as a source for automatic indexing. If we select the wrong preferred term, we might for example end up suggesting to index a document about food with the term fertilizer, because the non preferred term minerals has both food and fertilizer as its preferred term. We will present in the next section our semi-automatic annotation pipeline, the ranking algorithm applied to the term extraction and its usefulness for selecting the right preferred term out of 2 to 3 different possibilities.

3.4 Semi-automatic annotation pipeline

3.4.1 The pipeline

As stated in section 2, the goal of the semi automatic annotation pipeline is to suggest appropriate indexing terms to cataloguers with the goal of easing their job and increasing productivity. From discussion with the cataloguers it followed that they like a focussed and limited set of keywords : focussed because they only experience a suggestion as supportive if it closely matches the main topic of the document, limited because actual work process of cataloguers only allows for a limited number of terms to be attached to a document and because the inspection of the suggested terms should improve the work process, so the inspection time and the mental processing of the suggestions need to be bounded.

The pipeline consists of tree parts : a term detector, a term collector and a term ranker. As input to our pipeline we use our selected corpus and the GTAA. The output of the pipeline is a ranked list of GTAA preferred terms.

3.4.2 The input : GTAA in a RDF-OWL representation

As input we use an RDF-OWL representation of the GTAA, based on the SKOS Working Draft (see (van Assem *et al.*, 2006) and (Miles & Brickley, 2005)). The SKOS representation of a thesaurus is "concept based" : instead of terms, the entities are nodes with identifiers (ID), to which

[4] All the terms we mention in this paper are translated from Dutch to English out of consideration for our readers. We tried to select examples which have the same ambiguity in their semantics in the English translation

labels are attached, a prefLabel to represent the preferred term, and one or more altLabel(s) to represent the non preferred term(s). As the GTAA entries are in plural form, we also extended this model to add the information of the singular form corresponding to the original thesaurus terms. This model has drawbacks, and has an obvious conceptual bias, but it helps gathering pragmatically different strings corresponding to the same annotation ID. These strings are called "textual representations of the concept" in the GATE pipeline, and we decided to keep this terminology here.

3.4.3 The term detector : GATE with the Apolda plug-in

The term detector scans a text and looks for all possible textual representations of concepts. The detector is built with the Apolda plug-in in GATE architecture (Maynard *et al.*, 2003). After tokenization, the Apolda plug-in makes a simple string matching. It annotates a piece of text with the ID of the "concept" corresponding to the longest matching textual representation. If for a piece of text multiple concepts have the same longest matching textual representation, which can be the case for a non preferred term with multiple preferred terms, the plug in generates all possible annotations. This means that the string minority will receive two annotations : Keyword_ethnical_minority and Keyword_religious_minority. The string religious minority however will only receive the latter. The term detector is not case sensitive.

3.4.4 The term collector

The outcome of the term detector is an annotated text. In this text, multiple annotations can correspond to the same "concept". The term collector collects all the annotation ID's, computes their number of occurrences and writes the output into one file.

3.5 The term ranker : the CARROT algorithm

The file with ID's and number of occurrences computed at the previous step is fed into the Cluster And Rank Related to Ontology and Thesauri algorithm (CARROT algorithm). The CARROT algorithm is described in the section ranking algorithm of (Gazendam *et al.*, 2006).

CARROT uses the fact that terms in the Keyword facet of the GTAA are related to others via the related term, broader term and narrower term relations. Terms which relate to a lot of the other found terms in the text can semantically be more representative of the text than terms which are found more often but without any relations to others. If one of the thesaurus relationship exists between two of the found terms we say that a relation of distance 1 exists. We also check if an intermediate term connects two terms in the GTAA. These connections via intermediate terms are defined as relations of distance 2. We do not make any distinction in the type of relationships.

To make a ranking from the found keywords, we use the following rules :

- Step 1. We select the keywords with both a distance 1 and a distance 2 relation. We then order these keywords based on their number of occurrences, putting the most frequent on top of the list.
- Step 2. We select the remaining keywords with a distance 2 relation to keywords found during Step 1. We order these keywords based on their number of occurrences and add them to the

list.

- Step 3. We select the remaining keywords with a relation. We order these keywords based on their number of occurrences and add them to the list.
- Step 4. We order the remaining keywords based on their number of occurrences and add them to the list.

This algorithm creates clusters of ranked terms (several terms can have the same rank, they are then simply ordered alphabetically). Our previous experiment in (Gazendam *et al.*, 2006) showed that only the top clusters provided relevant keywords, so we intend to present the cataloguers with only these top clusters by default, with the possibility to access the whole ranked list if they wish to. In this paper we propose this CARROT algorithm as a means for selecting the right preferred term (right interpretation) for a non preferred term with multiple preferred terms (an ambiguous word). As the non preferred term attributes the same number of occurrences to all its preferred terms, three scenarios are possible :

- one of the preferred terms has more direct or indirect relations to other found terms and ranks higher as a result
- One of the preferred terms gets more occurrences due to the fact that the preferred term appeared itself in the text or one of its other non preferred terms appeared in the text
- both preferred terms rank equally high

So as an example let us run the following text through our pipeline :" *Snacks do not contain a lot of minerals.*" the non preferred term minerals has three preferred terms : food, fertilizer and ore. All are considered to occur once, because their common non preferred term occurs once. Another found term is snacks. These four terms are fed into CARROT. Due to the direct relation between food and the found term snacks, food now ranks higher the the other two preferred terms. This means that we here interpret minerals as referring to food.

As the output of the pipeline is the same list of annotation ID's as the input, but ranked, our hypothesis for disambiguating this case of polysemy is that the irrelevant preferred terms will not be connected to any of the other found keywords, and thus will be ranked at the bottom of the list, and, as a consequence, not shown to the cataloguers as indexing suggestion. We present the positioning of our experiment with the state of the art in Word Sense Disambiguation in the following section, followed by the experiment itself.

4 Related Work

The task we are interested in in this paper can be related to Word Sense Disambiguation. In (Ide & Véronis, 1998), the authors describe the typical two-step process for this task :

1. Define the set of senses per lexical unit ;
2. Use either a context-based method to determine which of the senses corresponds to the occurrence of the lexical unit considered, or an external knowledge source.

Many works mention the use of a dictionary as an external knowledge for that purpose ((Veronis & Ide, 1990), for example), whereas statistically-based or machine-learning methods advertise the corpus-based contextual approach (see for example (Yarowsky, 1995)). Of course, some mixed approaches exist, as (Stevenson & Wilks, 2001). In our use case, the set of senses to take into account in the set of possible preferred terms for each ambiguous non preferred term. The method that we experiment here is using external knowledge, but instead of the lexical content of dictionary definitions, or instead of trying to map lexical environment of the external knowledge to the corpus content, we use the thesaurus independently, and take only into account the number of occurrences of each term as a contextual information. The selection of the relevant sense, *i.e.* of the relevant preferred term, is made only based on relationships crafted by hand by cataloguing experts when building the thesaurus. Therefore it is still different from (Yarowsky, 1992), who also based his Word Sense Disambiguation algorithm on a thesaurus.

5 The experiment

5.1 Experiment : selecting the right keyword when multiple USE relations are possible

For this experiment, we annotated our documents with all the possible preferred terms related to the non preferred terms we found in the texts, along with the occurrences of the preferred terms found, and we will check whether the algorithm designed for ranking the output will help us disambiguating between the different possibilities, and :

1. Rank higher the relevant preferred term ;
2. Rank the irrelevant preferred terms low enough for them not to be part of the keywords suggestions made to the cataloguers.

5.2 Material

We constructed our corpus from a set of over 500 catalogue descriptions from Sound and Vision, related to TV programs. Each of these catalogue descriptions contains specific fields, that are described in Dublin Core : *e.g.* maker, title and keywords. One of the fields is a free text description called summary. In the keyword field the topic of the program is described by a limited set of preferred terms from GTAA's keyword facet. From this set of catalogue descriptions we selected all files which :

1. contain a non preferred terms which has multiple preferred terms and
2. have one of its related preferred terms appear in the keyword field

Based on these requirements we selected automatically a set of 121 documents, constituting our corpus. Each document contains in average 200 words. The requirement to have one of the preferred term appear in the keyword field gives us an evaluation criterium : we know that the document can semantically be described with the preferred term, so this preferred term can be seen as the correct interpretation of the non preferred term.

5.3 Experiment

We ran our pipeline on our corpus. After completion we looked at the non preferred term, the rank of all associated preferred terms on the ranked list and compared this ranked list with the preferred term in the keyword field of the catalogue description. We have four possible outcomes of this comparison :

1. The data is unusable
2. Correct suggestion : the suggested preferred term[5] is the preferred term in the keywords

[5] *i.e.* the preferred term with the highest rank in the list.

3. Wrong suggestion : the suggested preferred term is not the preferred term in the keywords
4. Undecidable : No suggestion is made because two (or all three) preferred terms rank equally high

The results are in table 1

correct	undecidable	wrong	unusable data	total
43	26	2	50	121

TAB. 1 – Results

5.3.1 Discussion

One of the first thing that seems remarkable is that we still have a lot of unusable data in our corpus. We find three types of such data :

- Documents which do not contain an ambiguous non preferred term.
- Documents which do not contain a preferred term related to an ambiguous non preferred term in the keyword field.[6]
- Documents in which the non preferred term is found in the keyword field. According to the production rules of Sound and Vision no non preferred term can be used in the keyword field. The set of keywords however changes over time. A preferred term may be ambiguous and as a consequence be changed to a non preferred term. Because we used older descriptions in our corpus, some of these contain previously preferred terms which now are non preferred terms in their keyword field. This means that they do not qualify as good documents for our corpus and are deselected. Examples are moordaanslagen and tentenkampen. These two examples account for two thirds of the unusable data : respectively 8 and 23 occurrences.

We excluded these from our analysis.

For the remainder of the corpus, in approximately 19 out of 20 cases, the suggestions are correct. We found only two cases in which we gave a wrong suggestion. Both mistakes are with the same non preferred term clubs which has as preferred terms hotel, restaurant and cafe and association. This word club was used in the context of football clubs. One text was on the share issue of soccer club Ajax. The other text was on the showing of a documentary on the soccer club Ajax in a theater. The term club had the meaning of association in both cases, referring to the soccer association. However the hotel, restaurant and cafe was suggested. In both cases terms at distance 2 from hotel, restaurant and cafe where present in the text : theater via the intermediate term nightlife and director via the intermediate term enterprize. On the other hand associations did not give connections to other found terms in the football domain as soccer, supporter, match, trainer because the distance was too big. Of the two other cases with the club the matching to its preferred terms is successful once and undecidable another time. Both these texts where also in the soccer domain and having the preferred term association. This could give the suggestion that elements in our corpus usually have one "preferential preferred term". This is not the case : the non preferred term windmills occurs once as wind turbine and once as mill. In both cases the correct suggestion is made, other non preferred

[6] We use the presence of one of the preferred terms in the keyword field to determine the correct one, so with no relevant preferred term among the keywords, we cannot evaluate our algorithm. These two cases result from errors in selecting documents for the corpus.

terms with a bigger number of occurrences have a non regular distribution of their preferred terms.

Another remarkable feature of the method is the big number of undecidable cases. The reason why we encounter this big number of undecidable cases is manyfold :

1. Our method uses general conditional rules. These conditions are not really specific : *having any distance 1 relation satisfies a condition*. As a result, in many cases both preferred terms fit the same conditions. This can be amended by sharpening these conditions, for example by counting the number of terms at distance 1 or distance 2.

2. The texts of our corpus are relatively small, so the number of found (and related) terms is also small, and the number of occurrences too low to disambiguate between different possibilities.

3. In many cases both preferred terms have a distance 1 relation. This means that already one of the few conditions for discrimination is ruled out, increasing the chance of a tie. At the same time this means that the difference in meaning between the preferred terms is subtle, making the fact that the algorithm cannot decide not necessary a bad property (could you decide between the meanings of toxin meaning poison, venom and dangerous substance in a text on farmers getting ill after using a toxin as a form of herbicide ?).

The last remark that we can make is that, due to the small number of occurrences of each of the different keywords in the different texts, very few clusters were created. As a consequence, it was hardly ever the case that the non relevant preferred terms found place low enough in the ranked list not to be proposed for indexing suggestion. Therefore, we should modify our algorithm in order to make it take into account only the preferred term with higher rank, and remove the other related preferred terms, in order not to generate noise in suggesting metadata to the users.

6 Conclusion and Perspectives

We investigated wether our method with the CARROT algorithm could be used for disambiguation in an indexing setting. It only gives a suggestion which preferred term to use for a non preferred term for 2 in 3 cases, but when it gives a suggestion, it does so in approximately 19 out of 20 cases correctly. It even showed that the bad suggestion came from one specific instance of the 83 non preferred terms. For this term we might include a *cave at* in our method. The structure of the thesaurus on the one hand and the use of terms associated to the wrong preferred term in this football club domain can result in these errors. If we incorporate this *cave at* we might come close to our goal : give good suggestions to cataloguers. The interpretation of our success rate and percentage of undecidable cases however must be subject of study : the cataloguer judgement of the support given determines wether these numbers are fair [7]. This will be subject of another study.

[7] A success of 19 out of 20 seems quite reasonable in the perspective of IR publications, but when talking about automatically securing railway crossings, the same success ratio is considered really bad.

Acknowledgements

This research was partly supported by the NWO funded CATCH program, including the CHOICE project. We want to thank our colleagues, both in Academy and at Sound and Vision for they daily help, support and our fruitful collaboration.

Références

GAZENDAM L., MALAISÉ V., SCHREIBER G. & BRUGMAN H. (2006). Deriving semantic annotations of an audiovisual program from contextual texts. In *Proceedings of First International workshop on Semantic Web Annotations for Multimedia (SWAMM 2006)*.

IDE N. & VÉRONIS J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Computational Linguistics*, **24**(1), 1–40.

ISO (1986). Documentation - guidelines for the establishment and development of monolingual thesauri. Iso 2788-1986.

MAYNARD D., TABLAN V. & CUNNINGHAM H. (2003). Ne recognition without training data on a language you don't speak. *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition : Combining Statistical and Symbolic Models*.

MILES A. & BRICKLEY D. (2005). Skos core guide. 2nd W3C Public Working Draft.

STEVENSON M. & WILKS Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, **27**(3), 321–349.

VAN ASSEM M., MALAISÉ V., MILES A. & SCHREIBER G. (2006). A method to convert thesauri to skos. In *Proceedings of the Third European Semantic Web Conference (ESWC'06)*.

VERONIS J. & IDE N. M. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics*, p. 389–394, Morristown, NJ, USA : Association for Computational Linguistics.

YAROWSKY D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of 14th International Conference on Computational Linguistics (COLING-92)*.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics (ACL' 95)*.