SKELETON: Specialised knowledge retrieval on the basis of terms and conceptual relations

Judit Feliu, Jorge Vivaldi, M. Teresa Cabré

Institute for Applied Linguistics, Universitat Pompeu Fabra
La Rambla 30-32, 08002 Barcelona, Spain
{judit.feliu;jorge.vivaldi;teresa.cabre}@upf.edu

Abstract

The main goal of this paper is to present a first approach to an automatic detection of conceptual relations between two terms in specialised written text. Previous experiments on the basis of the manual analysis lead the authors to implement an automatic query strategy combining the term candidates proposed by an extractor together with a list of verbal syntactic patterns used for the relations refinement. Next step on the research will be the integration of the results into the term extractor in order to attain more restrictive pieces of information directly reused for the ontology building task.

1 Introduction

In this paper the authors show a strategy planned to obtain specialised knowledge fragments containing terms together with the conceptual relations among them, that is, the skeleton of a text that could be schematised by means of a concept map and hopefully reused in order to enrich a domain dependant ontology. These terms and relations will be detected in written texts from the genomics domain. The results presented in this paper have been obtained for the Catalan language but we are already working on the implementation of the same working methodology for specialised texts in Spanish and English will be also considered in a near future.

Roughly speaking, our proposal shows one of the methodologies used for the achievement of conceptual mapping from texts and it includes two different and complementary strategies:

On the one hand, we have used a term extractor (YATE) in order to obtain the term candidates in genomics domain texts. YATE has been tuned to cover the working specialised field by means of the enlargement and refinement of some domain dependant information. And, at the same time, the improvement of YATE has contributed to the enlargement of EuroWordNet (a wide-coverage general-purpose lexico-semantic ontology) with new synsets.

On the other hand, we have reviewed the traditional conceptual relations classification from the point of view of different (but closely related) disciplines, such as terminology, linguistics, ontologies and lexical semantics. After a validated experiment, we have proposed a closed typology of conceptual relations including seven main types of links that may relate the terms used in any domain, therefore also in genomics. These conceptual relations are reflected, in terms of language, by means of verbal markers usually accompanied with prepositions among other language specific mechanisms not used in our experiments. This patterns have been applied in order to compare the information contained between two different terms and to tag specialised knowledge fragments.

In this paper after a brief state-of-the-art about conceptual relations, and the automatic detection strategies of these links, it is included the preliminary analysis of the verbal markers concerning precision and noise. Manually detected patterns from a sample corpus have allowed the authors to explore and implement an automatic query system which has been progressively refined. Some illustrating and relevant contexts are highlighted in the results section indicating some figures concerning precision for each verbal pattern conveying a particular conceptual relation. It is worth mentioning that the integration into YATE of the obtained results using a kwic query interface is described in the future research lines before briefly concluding the paper.

2 State of the art

This section includes a brief state-of-the-art about conceptual relations, focusing on the point of view of the terminology contributions in this issue, and also some comments on recent advances in the automatic detection of some of these types of links.

2.1 Terminology

Concepts and conceptual relations are essential items of the cognitive structure concerning specialised knowledge. In specialised texts, concepts are expressed by terms and conceptual relations become the so-called semantic relations when trying to identify them using different linguistic expressions It is assumed that each terminological unit corresponds to a cognitive node in a specialised domain and the whole of these nodes are linked by means of particular conceptual relations. The number and typology of conceptual relations considered by the traditional approach to terminology are not representative enough in order to account for the real links established among terms in a specialised text (Cabré, 1999; Feliu et al., 2002). The abstraction about this primary issue in Terminology proposed by Wüster and some of his followers (Felber, 1984; Felber et al., 1984) becomes excessively restrictive in order to give account of the real connections found in specialised texts. Although the hierarchical and non-hierarchical taxonomy of conceptual relations (followed by ISO standards) has been useful in the classical terminographical vocabularies, later research shows the need for an enlargement of the number and the subespecification of conceptual relations. Just to mention a few, there have been interesting contributions on concrete types of relations (Winston et al., for the meronymy relation; Marshman et al., in the case of causality). After having reviewed and compared these

proposals with the typology of relations included in different well-known ontologies (EWN, MikroKosmos among others) we have set a clear-cut definition of conceptual relation and we have proposed an integrating typology of this key element in terminology.

Following Otman (1996), we define a conceptual relation as a binary link, between at least two terms, which has a semantic content itself; in other words, that it transfers a particular information and it allows to configure a specific predication among the terms it relates. Considering the potential recursivity of a conceptual relation, it can be represented by the following formula: <a R b, n>. The variables *a* and *b* represent the two minimum terms that have to appear between a conceptual relation, *R* represents the relation expressed by a linguistic marker and *n* foresees the possibility of some other terms to appear linked by that conceptual relation. Notice the following example for illustrating the above mentioned definition of conceptual relation:

34% of [human genes]$_a$ ***[are located in]***$_R$ *[L1]*$_b$ *and [L2 isochores]*$_n$.

Where:

a: "human genes"

R: "are located in" (sequenciality-place)

b: "L1 (isochore)"

n: "L2 isochore".

As already mentioned, the traditional typology of conceptual relations repetitively used in terminology has to be reconsidered when working with corpora. Our theoretical review, together with two experiments on specialized texts on medicine and genetics, carried us to set the following typology of conceptual relations[1] (Feliu, 2004):

a) Similarity

positive: *ser semblant a* [be similar to].

negative: *ser diferent de* [be different from].

b) Inclusion

Class inclusion or hyponimy (is-a relation): *ser (un tipus) de* [to be (a kind of)/is a].

c) Sequenciality

place: *ser en / ser davant / ser darrere / anar de x a y* [to be in/in front of/behind, to go from *x* to *y*].

time: *ser simultani / anterior / posterior a* [to be simultaneous/previous/later to].

d) Causality

causal: *causar / ser la causa de / ser l'efecte de / produir / fer que* [to cause/to be the cause of/to be the effect of/produce/make].

e) Instrument: *servir per a / fer-se amb* [to be useful for/to use/to be done with].

f) Meronymy: *ser una part / element de; tenir + SN / estar format / fet per; incloure; constar de / pertànyer a* [to be part/element of; have + SN, to be constituted of; to include]

g) Association

general: *correlacionar-se amb*; *mostrar* [to be related with / to show].

specialised: *manifestar; determinar* [to manifest / to determine].

This classification will become one of the essential items in the working methodology followed in order to attain the main goal of this research.

2.2 Automatic learning of semantic relations

There is a number of applications in NLP where conceptual hierarchies are used as a background knowledge for carrying out a given task. These resources are important because they allow to structure information into categories, thus encouraging its search and reuse. Further, they allow to formulate rules as well as relations in an abstract and concise way, facilitating the development, refinement and reuse of a knowledge base. Further, the fact that they allow to generalize over words has shown to provide benefits in a number of applications[2] such as text classification (Bloehdorn et al., 2004) word sense disambiguation (Agirre et al., 2004) and term extraction (Vivaldi et al., 2002).

However, it is well known that the development of an ontology represents a major bottleneck for any of the above mentioned NLP applications. Such difficulty arises because current methods for compiling an ontology relies on a manual enumeration of concepts and relations resulting in a high resources consuming task. Therefore, some effort has been done to overcome this issue by automatically acquiring ontological knowledge from domain-specific natural language texts. Most of the researchers in this area only considered to learn taxonomic relations. To mention but a few, we refer to some fairly recent work, e.g., by Hahn & Schnattinger (Hahn & Schnattinger, 1998) and Morin (Morin, 1999) who used lexico-syntactic patterns with and without background knowledge, respectively, in order to acquire taxonomic knowledge. Although taxonomic relations are of major significance some effort must be done to non taxonomic relations also. As a matter of fact, this is an issue not considered since recently. As an example, the user can /query an important ontology like EuroWordNet (or WordNet, its counterpart) to check that these kind of relationships are missing.

A basic and largely exploited idea is the heuristic, reported in Hearst (1992), that certain patterns in texts induce a hyponym relation between words. The following are some examples of such patterns:

- NP_0 such as $\{NP_n,\}_{n>0}$* (or | and) NP
- NP {, NP}*{,} or other NP
- etc.

An interesting approach that combine different sources of evidences is presented in Cimiano et al. (2004). Here the authors basically apply the Hearst patterns to a corpus, the World Wide Web, WordNet and knowledge obtained from complex terms.

A complete platform for semiautomatic acquisition of conceptual relations is proposed in Maedche et al. (2000). The authors perform an extensive and detailed linguistic analysis to the text chosen by an ontology engineer to learn from. Then the system applies a learning algorithm to find association rules among pairs of concepts. This system was tested in the tourism domain.

[1] It is worth mentioning that, as a working methodology, we established, for each conceptual relation, its corresponding prototype linguistic expression in order to tag the relations appearing in our corpus [an approximate Englisht translation is provided in square brackets].

[2] In http://mira.csci.unt.edu/~wordnet/ it is possible to check an extensive bibliography about WordNet and its applications to many areas of NLP.

In Missikoff et al. (2002), the authors present a software environment that supports the construction and assessment of a domain ontology.

3 Previous work

Concerning the experiment, a corpus on the human genome domain containing about 100.000 words from 18 different documents in Catalan taken from the IULA's technical corpus was used as an initial workbench. We have applied to the corpus a list of 55 verbal markers (manually validated in a previous research) using a term detector called *Mercedes.* This tool works on the basis of the comparison of the terms appearing in the texts with a set of terminological units contained in a predefined dictionary (see Araya et al, 2004, for details).

The integration of the list of verbs to the terminological units detecting system have resulted into more than three thousand contexts including, in the sentence framework, at least one terminological unit and the verb itself which is potentially expressing a conceptual relation. Next step on the corpus analysis has consisted in the manual validation of the retrieved contexts. Figures on relevance, precision and noise are shown on Table 1.

Verbal marker	Occ.	Precision	Noise
allunyar [to put far away]	4	75%	25%
aparèixer [to appear]	20	30%	70%
apropar [to put near]	1	0%	100%
arribar [to arrive]	26	15,38%	84,62%
caracteritzar [to characterise]	33	66,7%	33,3%
causar [to cause]	44	84,1%	15,9%
compondre [to compose]	5	20%	80%
considerar [to consider]	23	21,74%	78,26%
constar [to form part]	9	33,3%	66,7%
constituir [to constitute]	27	70,37%	29,63%
continuar [to follow]	10	20%	80%
contribuir [to contribute]	4	75%	25%
correlacionar [to correlate]	1	0%	100%
correspondre [to correspond]	25	68%	32%
definir [to define]	15	20%	80%
dependre [to depend	18	66,7%	33,3%
determinar [to determine]	58	72,41%	27,59%
deure [to due]	72	84,72%	15,28%
diferenciar [to differentiate]	16	62,5%	37,5%
distinguir [to distinguih]	8	50%	50%
donar [to give]	88	34,1%	65,9%
englobar [to include]	5	100%	0%
evidenciar [to make evident]	9	66,7%	33,3%
fer [to make]	225	8,4%	91,6%
formar [to form]	107	68,22%	31,78%
implicar [to imply]	49	71,43%	28,57%
incloure [to include]	33	90,90%	9,10%
indicar [to indicate]	52	75%	25%
iniciar [to initiate]	17	41,18%	58,82%
integrar [to integrate]	15	40%	60%
intervenir [to take part in]	14	71,43%	28,57%
localitzar [to localise]	26	61,54%	17,86%
manifestar [to manifest]	16	68,75%	31,25%
mesurar [to measure]	7	0%	100%
mostrar [to show]	66	59,1%	40,9%
originar [to originate]	20	50%	50%
presentar [to have]	98	72,45%	27,55%
produir [to produce]	111	42,34%	57,66%
propagar [to propagate]	1	0%	100%
provocar [to provoke]	35	68,57%	31,43%
quedar [to stay]	27	14,81%	85,19%
realitzar [to carry out]	43	25,58%	74,42%
reflectir [to reflect]	2	50%	50%
representar [to represent]	39	69,23%	30,76%
reunir [to gather together]	3	0%	100%
ser [to be]	1.045	33,11%	66,89%
simular [to simulate]	1	100%	0%
situar [to place]	29	75,86%	24,14%
suggerir [to suggest]	15	6,7%	93,3%
tenir [to have]	240	58,75%	41,25%
transcórrer [to take place]	2	0%	100%
trobar [to find]	124	27,42%	72,58%
usar [to use]	1	0%	100%
utilitzar [to use]	82	45,12%	54,88%
veure [to see]	48	4,17%	95,83%

Table 1. Verbal markers with frequency and precision indication

These figures must be considered being aware that all of the verbal markers show a different percentage of appearance. One of the main conclusions of the experiment led us to improve the precision ratio with the use of syntactic patterns. It is foreseen that these patterns will be helpful in the disambiguation process for those cases where the detection of the verb alone had produced polysemous conceptual relation tagging. Table 2 shows a sample of the combination of verb and a prepositional group.

Verb	Prepositions	Relation
allunyar	com	neg. sim.
aparèixer	a / en	seq. place loc.
arribar	a	seq. place dir.
caracteritzar	per	gen. ass.
constar	de	mer.
continuar	en	seq. temp. ant-pos.
definir	amb / per	inst.
diferenciar	cap a / de / en	spec. ass.
diferenciar	de / per	neg. sim.
donar	lloc a / origen a	caus.
estar situar	per	seq. temp. sim
incloure	a / en / entre	incl.
incloure	a / en	mer.
ser	un / el	incl. (is-a)
...	..	…

Table 2. Sample of verbs and their syntactic patterns

Notice that some verbs plus prepositional group may indicate two different relations (*diferenciar de* → specialised association or negative similarity), this results into an ambiguity.

4 Working methodology

Following the experiments and their results presented in the previous section, it was necessary to broad the

experiments with whole patterns, that is verb plus preposition or prepositional group (single verbs have been left aside). Additionally it was decided to use YATE in order to facilitate the validation process.
For such purpose we have designed a system that has the architecture shown Figure 1. This system, applies the term extractor to the same texts used in the previous experiment and then, the resulting terms are used to query the full genomic corpus looking for sentences where a pattern like <term><verb+prep> applies. The second part of the pattern corresponds to the sequences containing the verb plus preposition pre-defined in section 3.

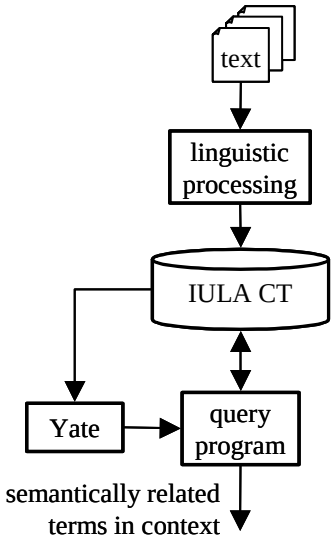

Figure 1. Query environment

In this context, the idea is to extract from the corpus only those sentences that are the best candidates to provide interesting information regarding the concepts transfered by means of the terms appearing in the texts.
Most of the systems already proposed for a similar task have a simplistic strategy for detecting the terms present in a given text. Here we propose to use YATE, a hybrid term extractor developed for the Medical domain that combines the results obtained by a set of term analyzers described briefly as follows (see Vivaldi, 2001 for details):

- domain coefficient: it uses the EWN ontology to sort the term candidate.
- context: it evaluates each candidate using other candidates present in its sentence context.
- classic forms: it tries to decompose the lexical units in their formants, taking into account the formal characteristics of many terms in the domain.
- collocational method: it evaluates multiword candidates according to its mutual information.

The results obtained by this set of heterogeneous methods are combined using two different methods: voting and boosting. In the former each single term analyzer reports a term/no term status while the latter makes use of a well-known method originated in the machine learning area.

5 Results

The results obtained from applying the above mentioned methodology are depicted in Table 3. It shows that the number of the specialized knowledge fragments retained by the resource is quite impressive. In terms of system evaluation, figures indicate a (relatively) high precision concerning the fine tagging of the contexts containing clear terms and one single type of conceptual relation.
Conversely, nothing can be said about the recall because, as known, in order to calculate this figure it should be necessary to know in advance all the relations included in the texts.

Relation	#	Valid	%
spec. ass.	53	50	94.34
spec. ass. / neg. sim.	5	3	60.00
spec. ass./ seq. place. loc.	37	37	100.00
gen. ass.	108	82	75.93
gen. ass. / seq. place. loc,	2	0	0.00
caus.	187	169	90.37
incl. / mer.	3	3	100.00
inst.	19	13	68.42
mer.	43	42	97.67
neg. sim.	3	3	100.00
pos. sim.	109	40	36.70
seq. place. dir.	5	2	40.00
seq. place. loc	157	119	75.80
incl.	383	147	38.38

Table 3. Obtained results

A close examination of the results in the previous table shows that for some non-ambiguous relations (general association, specialised association, causality, instrumentality, meronymy and sequenciality-place loc.) the performance reached is pretty good. On the contrary, some other relations like inclusion/meronymy and specialised association/sequenciality place loc., are ambiguous and therefore they require manual validation. In some cases, the number of textual fragments retained is too low and for this reason the results concerning these cases are not significant.
A separate comment on the inclusion relation expressed by the is-a marker should be done. In this case, the precision is quite low (nearly 40%) mainly due to the fact that the second term in the relation is not detected. When this position is occupied by words expressing exemplification, case or results among others, the inclusion and the localization relations do not apply. Also, when the word expresses cause or consequence the meaning of the marker itself changes to causality.
Conversely, the best results have been obtained with the causality and meronymy relations (about 90% and 97% of precision, respectively), although the former is much more frequent. The following two fragments show examples where these relations are correctly detected:

- causality: La #malaltia# de Gaucher és un trastorn degut a la disfunció del sistema lisosòmic. [Gaucher's #disease# is a change due to the lisosomic system disfunction]

- meronymy: Cada #gen# està format per una seqüència concreta de nucleòtids. [Each #gene# is formed by a concrete sequence of nucleotides]

Some of the non-retained fragments include the verbal pattern but in some cases, the term detected is not the

subject of the verb expressing the relation and, in some others, the textual fragment appearing after the prepositional group does not include a term. The following example show these two inconveniences at the same time:

- El coneixement de l'estructura tridimensional de les #molècules# de classe I és degut als treballs d'estructura cristal·logràfica de les molècules HLA-A2, HLA-A68i HLA-B27. [The study of the tridimensional structure for type I #molecules# is due to the research works on the structure…]

The next fragments show some examples where the is-a relation is correctly detected:

- L'estrès oxidatiu ha estat implicat en nombroses #malalties neurodegeneratives humanes# com són la #SD#, les malalties d'Alzheimer, Parkinson i Huntington, l'esclerosi lateral amiotròfica (als), distròfies musculars, arteriosclerosi i els processos d'envelliment en general. [... human neurodegenerative diseases such as DS, Alzheimer's diseases ...]
- Per fi, després de gairebé un segle d'història, el #cromosoma#, que és una #molècula contínua d'ADN#, es pot parcel·lar en les seues unitats discretes, els gens. [... the chromosome is a continuous DNA molecule ...]
- La #proteïna HMG-14# és una #proteïna nuclear# que podria modular l'estructura de la cromatina transcripcionalment activa. [The HMG-14 protein is a nuclear protein ...]

The following example is not valid due to the anaphoric reference present after the pattern instead of the desired term:

- Una de les tècniques més prometedores per definir xarxes de #gens# entrellaçats és la dels denominats bioxips .

6 Future work

In order to facilitate the building task to the ontology engineer we foresee a tight integration of YATE in such way that the proposed system will be looking for the triplet (term1, relation, term2). On the basis of the information given regarding Figure 1, this improvement means that the extractor itself would be able to look for the sentences containing the highest ranked terms and only for these sentences it will try to find a valid semantic relation indicator.

In order to confirm these promising results we plan to apply the same procedures to other domains, as wells as to other languages like Spanish and afterwards English.

7 Conclusions

We have presented a research based on learning semantic relations from text in the domain of genomics. In particular, we investigate the use of verbs as semantic relations markers. Our model is based upon a previous work on the definition of a typology of conceptual relations and some experimentation using this set of markers. As a result of this experimentation it was decided to attach a specific preposition to every verb in order to refine the results and to avoid ambiguity derived from polysemous markers. A new experiment was designed introducing the use of a term extractor together with a corpus query program. The result of this new experiment is very encouraging because an interesting number of text fragments can be considered as correctly detected concerning specialised knowledge retrieval. This result encourage us to go further in the integration of the proposed methodology with a term extractor.

To sum up, we conclude that the specialised knowledge retrieval using terminological information in order to get a semiautomatic ontology and term database enlargement may be possible with the use of the techniques that we have presented in this paper.

8 References

Agirre E., Martinez, D. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Barcelona: Spain.

Araya, R.; Vivaldi, J. (2004). Mercedes, A Term-In-Context Highlighter. In: *Proceedings of Language and Resources Evaluation Conference (LREC2004).* Lisbon.

Bloehdorn, S., Hotho, A. (2004). Boosting for Text Classification with Semantic Features. In *Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 70-87.

Cabré, M.T. (1999). La terminología: Representación y comunicación. Una propuesta de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cimiano, P.; Pivk, A.; Schmidt-Thiene, L.; Staab S. (2004). Learning Taxonomic Relations from Heterogeneous Evidence. In *ECAI-2004 Workshop on Ontology Learning and Population*.

Feliu, J. (2000). Relacions conceptuals i variació funcional: elements per a un sistema de detecció automàtica. Barcelona: Institut Universitari de Lingüística Aplicada [non-published].

Feliu, J. (2004). Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. Barcelona: Institut Universitari de Lingüística Aplicada. (Sèrie tesis, 7) [PhD dissertation published in a CD-ROM].

Feliu, J.; Cabré, M.T. (2002). Conceptual relations in specialized texts: new typology and an extraction system proposal. In *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, Nancy, pp. 45-49.

Felber, H.; Picht, H. (1984). Métodos de terminografía y principios de investigación terminológica. Madrid: Instituto Miguel de Cervantes (CSIC).

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th*

International Conference on Computational Linguistics, pp. 539-545.

Iris, M.A.; Litowitz, B.E.; Evens, M. (1988). The problems of the part-whole relation. In: *Relational models of the lexicon. Representing knowledge in semantic networks*. Cambridge: Cambridge University Press, pp. 260-287.

Maedche, A.; Staab S. (2000). Discovering Conceptual Relations from Text. In *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, 2000*. IOS Press, Amsterdam.

Marshman, E.; Morgan, T.; Meyer, I. (2002). French patterns for expressing concept relations. In: *Terminology*, 8, 1, pp. 1-29.

Marshman, E. (2002). The Cause-Effect Relation in a Biopharmaceutical Corpus: English Knowedge Patterns. In *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, Nancy, pp. 89-94.

Missikoff, M.; Navigli, R.; Velardi P. (2002). Integrated Approach to Web Ontology Learning and Engineering. *IEEE Computer*, November 2002.

Morin, E. (1999). Automatic acquisition of semantic relations between terms from technical corpora. In *Proceedings International Congress on Terminology and Knowledge Engineering-TKE-99*.

Otman, G. (1996). *Les représentations sémantiques en terminologie*. París: Masson.

Vivaldi, J., (2001). Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas. PhD dissertation. Universitat Politècnica de Catalunya.

Vivaldi, J.; Rodríguez, H. (2002). Medical Term Extraction using the EWN ontology. In In *TKE 2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th August 2002*, Nancy.

Winston, M.E.; Chaffin, R.; Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11, pp. 417-444.