Automatic Discovery of Term Similarities Using Pattern Mining

Goran NENADIĆ, Irena SPASIĆ and Sophia ANANIADOU

Computer Science, University of Salford
Salford, M5 4WT, UK
{G.Nenadic, I.Spasic, S.Ananiadou}@salford.ac.uk

Abstract

Term recognition and clustering are key topics in automatic knowledge acquisition and text mining. In this paper we present a novel approach to the automatic discovery of term similarities, which serves as a basis for both classification and clustering of domain-specific concepts represented by terms. The method is based on automatic extraction of significant patterns in which terms tend to appear. The approach is domain independent: it needs no manual description of domain-specific features and it is based on knowledge-poor processing of specific term features. However, automatically collected patterns are domain specific and identify significant contexts in which terms are used. Beside features that represent contextual patterns, we use lexical and functional similarities between terms to define a combined similarity measure. The approach has been tested and evaluated in the domain of molecular biology, and preliminary results are presented.

Introduction

In a knowledge intensive discipline such as molecular biology, the vast and constantly increasing amount of information demands innovative techniques to gather and systematically structure knowledge, usually available only from text/document resources. In order to discover new knowledge, one has to identify main concepts, which are linguistically represented by domain specific terms (Maynard and Ananiadou (2000)). There is an increased amount of new terms that represent newly created concepts. Since existing term dictionaries usually do not meet the needs of specialists, automatic term extraction tools are indispensable for efficient term discovery and dynamic update of term dictionaries.

However, automatic term recognition (ATR) is not the ultimate aim: terms recognised should be related to existing knowledge and/or to each other. This entails the fact that terms should be classified or clustered so that semantically similar terms are grouped together. Classification and/or clustering of terms are indispensable for improving information extraction, knowledge acquisition, and document categorisation. Classification can also be used for efficient term management and populating and updating existing ontologies in a consistent manner. Both classification and clustering methods are built on top of a specific similarity measure. The notion of term similarity has been defined and considered in different ways: terms can have functional and/or structural similarities, though they can be correlated by different relationships (Grefenstette (1994), Maynard and Ananiadou (2000)). In this paper we suggest a novel, domain-independent method for the automatic discovery of term similarities, which can serve as a basis for both classification and clustering of terms. The method is mainly based on the automatic discovery of significant term features through pattern mining. Automatically collected patterns are domain dependent and they identify significant contexts in which terms tend to appear. In addition, the measure combines lexical and syntactical similarities between terms.

The paper is organised as follows. In Section 1 we overview term management approaches. Section 2 introduces the term similarity measure and Section 3 presents results and experiments.

1 Terminology Management

Since vast amount of knowledge still remains unexplored, several systems have been proposed to help scientists to acquire relevant knowledge from scientific literature. For example, GENIES (Friedman et al. (2001)) uses a semantic grammar

and substantial syntactic knowledge in order to extract comprehensive information about signal-transduction pathways. Some of the systems are terminology-based, since technical terms semantically characterise documents and therefore represent starting place for knowledge acquisition tasks. For example, Mima et al. (2002) introduce TIMS, a terminology-based knowledge acquisition system, which integrates automatic term recognition, term variation management, context-based automatic term clustering, ontology-based inference, and intelligent tag information retrieval. The system's aim is to provide efficient access and integration of heterogeneous biological textual data and databases.

There are numerous approaches to ATR. Some methods (Bourigault (1992), Ananiadou (1994)) rely purely on linguistic information, namely morpho-syntactic features of term candidates. Recently, hybrid approaches combining linguistic and statistical knowledge are becoming increasingly used (Frantzi et al. (2000), Nakagawa et al. (1998)).

There is a range of clustering and classification approaches that are based on statistical measures of word co-occurrences (e.g. Ushioda (1996)), or syntactic information derived from corpora (e.g. Grefenstette (1994)). However, few of them deal with term clustering: Maynard and Ananiadou (2000) present a method that uses manually defined semantic frames for specific classes, Hatzivassiloglou et al. (2001) use machine learning techniques to disambiguate names of proteins, genes and RNAs, while Friedman et al. (2001) describe extraction of specific molecular pathways from journal articles.

In our previous work, an integrated knowledge mining system in the domain of molecular biology, ATRACT, has been developed (Mima et al. (2001)). ATRACT (**A**utomatic **T**erm **R**ecognition and **C**lustering for **T**erms) is a part of the ongoing BioPath[1] project, and its main aim is to facilitate an efficient expert-computer interaction during term-based knowledge acquisition. Term management is based on integration of automatic term recognition and automatic term clustering (ATC). ATR is based on the *C/NC-value* method (Frantzi et al. (2000)), a hybrid approach combining *linguistic knowledge* (term formation patterns) and *statistical knowledge* (term length, frequency of occurrence, etc). The extension of the method handles orthographic, morphological and syntactic term variants and acronym recognition as an integral part of the ATR process (Nenadić et al. (2002a)), providing that all term occurrences of a term are considered. The ATC method is based on the Ushioda's AMI (Average Mutual Information) hierarchical clustering method (Ushioda (1996)). Co-occurrence based term similarities are used as input, and a dendrogram of terms is generated.[2]

2 Term Similarity Measures

In this section we introduce a novel hybrid method to measure term similarity. Our method incorporates three types of similarity measures, namely contextual, lexical and syntactical similarity. We use a linear combination of the three similarities in order to estimate similarity between terms. In the following subsections we describe each of the three similarity measures.

2.1 Contextual Similarity

Determining the similarity of terms based on their contexts is a standard approach based on the hypothesis that similar terms tend to appear in similar contexts. Contextual similarity, however, may be determined in a number of ways depending on the way in which the context is defined. For example, some approaches consider only terms that appear in a close proximity to each other (Maynard and Ananiadou (2000)), while in other approaches, grammatical roles such as object or subject are taken into account (Grefenstette (1994)).

Our approach to contextual similarity is based on automatic *pattern mining*. The aim is to automatically identify and learn the most important context patterns in which terms appear. *Context pattern* (CP) is a generalised regular expression that corresponds to either left or right context of a term.[3] The following example shows a sample left context pattern of the term `high affinity`:

[1] BioPath is a Eureka funded project, coordinated by LION BioScience (http://www.lionbioscience.com) and funded by the German Ministry of Research.

[2] For the evaluation of the ATR and ATC methods incorporated in ATRACT, see Mima et al. (2001).

[3] Left and right contexts are treated separately.

V:*bind* TERM:*rxr_heterodimers* PREP:*with*

Let us now describe the process of constructing CPs and determining their importance. First, we collect concordances for all automatically recognised terms. Context constituents, which we consider important for discriminating terms (e.g. noun and verb phrases, prepositions, and terms themselves) are identified by a tagger and by appropriate local grammars, which define syntactic phrases (e.g. NPs, VPs). The grammatical and lexical information attached to the context constituents is used to construct CPs. In the simplest case, contexts are mapped into the syntactic categories of their constituents. However, the lemmatised form for each of the syntactic categories can be used as well. For example, when encountered in a context, the preposition `with` can be either mapped to its POS tag, i.e. `PREP`, or instead, the lemma can be added, in which case we have an *instantiated* chunk: `PREP:`*`with`*. Further, some of the syntactic categories can be removed from the context patterns, as not all syntactic categories are equally significant in providing useful contextual information (Maynard and Ananiadou (2000)). Such CPs will be regarded as *normalised* CPs. In our approach, one can define which categories to instantiate and which to remove. In the examples provided later in the paper (Section 3) we decided to remove the following categories: adjectives (that are not part of a term), adverbs, determiners and so-called linking words (e.g. `however`, `moreover`, etc.). Also, we instantiated terms and either verbs or prepositions, as these categories are significant for discriminating terms.

Once we have normalised CPs, we calculate the values of a measure called *CP-value* in order to estimate the importance of the CPs. CP-value is defined similarly to the C/NC-value for terms (Frantzi et al. (2000)). It assesses a CP (p) according to its total frequency ($f(p)$), its length ($|p|$, as the number of constituents) and the frequency of its occurrence within other CPs ($|T_p|$, where T_p is a set of all CPs that contain p):

$$CP(p) = \begin{cases} \log_2|p| \cdot f(p); & p \text{ is not nested} \\ \log_2|p| \cdot \left(f(p) - \frac{1}{|T_p|} \sum_{b \in Tp} f(b) \right) & \text{; otherwise} \end{cases}$$

The CPs whose CP-value is above a chosen threshold are deemed important. Note that these patterns are domain-specific and that they are automatically extracted from a domain specific corpus. Tables 1 and 2 show samples of significant left context patterns extracted from a MEDLINE corpus (MEDLINE (2002)).

CPs	**CP-value**
PREP NP	272.65
PREP NP PREP	186.47
. . .	. . .
PREP NP V:*stimulate*	9.32
V:*indicate* NP	5.00
PREP NP PREP V:*involve* NP	4.64
PREP TERM:*transcriptional_activity*	4.47
V:*require* NP PREP	4.38
PREP TERM:*nuclear_receptor* PREP	4.00

Table 1: Sample of left CPs
(only terms and most frequent verbs are instantiated)

CPs	**CP-value**
PREP:*of* NP	121.49
V NP	71.42
PREP:*of* NP V	62.83
NP PREP:*of* NP	59.72
PREP:*in* NP	59.55
NP PREP:*of*	43.37
PREP:*of* NP V NP	37.64
PREP:*of* TERM:*transcriptional_activity*	36.60

Table 2: Sample of left CPs
(only terms and prepositions are instantiated)

At this point, each term is associated with a set of the most characteristic patterns in which it occurs. We treat CPs as term features, and we use a feature contrast model (Santini and Jain (1999)) to calculate similarity between terms as a function of both common and distinctive features. Let us now formally define the contextual similarity measure. Let C_1 and C_2 be two sets of CPs associated with terms t_1 and t_2 respectively. Then, the *contextual similarity* (*CS*) between t_1 and t_2 corresponds to the ratio between the number of common and distinctive contexts:

$$CS(t_1, t_2) = \frac{2|C_1 \cap C_2|}{2|C_1 \cap C_2| + |C_1 \setminus C_2| + |C_1 \setminus C_2|}$$

2.2 Lexical Similarity

We also examine the lexical similarity between words that constitute terms. For example, if terms share the same head, they are assumed to have the

same concept as an (in)direct hypernym (e.g. progesterone receptor and oestrogen receptor). Further, if one of such terms has additional modifiers, this may indicate concept specialisation (e.g. nuclear receptor and orphan nuclear receptor). Bearing that in mind, we base the definition of lexical similarity on having a common head and/or modifier(s). Formally, if t_1 and t_2 are terms, H_1 and H_2 their heads, and M_1 and M_2 the sets of the stems of their modifiers, then their *lexical similarity* (*LS*) is calculated according to the following formula:

$$LS(t_1, t_2) = \frac{1}{a+b}\left(a * |H_1 \cap H_2| + b * \frac{2|M_1 \cap M_2|}{2|M_1 \cap M_2| + |M_1 \setminus M_2| + |M_1 \setminus M_2|}\right)$$

where a and b are weights such that $a > b$, since we give higher priority to shared heads over shared modifiers.

Note that the lexical similarity between two different terms can have a positive value only if at least one of them is a multiword term. Also, when calculating lexical similarity between terms that are represented by corresponding acronyms, we use normalised expanded forms. [4]

2.3 Syntactical Similarity

By analysing the distribution of similar terms in corpora, we observed that some general (i.e. domain independent) lexico-syntactic patterns indicate functional similarity between terms. For instance, the following example:

... steroid receptors *such as* estrogen receptor, glucocorticoid receptor, *and* progesterone receptor.

suggests that all the terms involved are highly correlated, since they appear in an enumeration (represented by the *such-as* pattern) which indicates their similarity (based on the *is_a* relationship). Some of these patterns have been previously used to discover hyponym relations between words (Hearst (1992)). We generalised the approach by taking into account patterns in which the terms are used *concurrently* within the same context. We hypothesise that the parallel usage of terms within the same context, as a specific type of co-occurrence, shows their functional similarity. Namely, all the terms within a parallel structure have the same *syntactic* function within the sentence (e.g. object or subject) and are used in combination with the same verb or preposition. This fact is used as an indicator of their semantic similarity.

In our approach, several types of lexico-syntactical patterns are considered: enumeration expressions, coordination, apposition, and anaphora. However, currently we do not discriminate between different similarity relationships among terms (which are represented by different patterns), but instead, we consider terms appearing in the same syntactical roles as highly semantically correlated.

A sample of enumeration patterns is shown in Table 3. [5] Manually defined patterns are applied as syntactic filters in order to retrieve sets of similar terms. These patterns provide relatively good recall and precision. We also used coordination patterns (Klavans et al. (1997)) as another type of parallel syntactic structure. Two types of argument coordination and two types of head coordination patterns were considered (see Table 4). However, not all the sequences that match the coordination patterns are coordinated structures (see Table 5). Therefore, these patterns provide relatively good recall, but not high precision if one wants to retrieve terms involved in such expression.[6] However, both term coordination and (nominal) conjunction of terms indicate their similarity.

Based on co-occurrence of terms in these parallel lexico-syntactical patterns, we define the *syntactical similarity* (*SS*) measure for a pair of terms as 1 if the two terms appear together in any of the patterns, and 0 otherwise.

[4] For our approach to acronym acquisition and term normalisation, see Nenadic et al. (2002).

[5] Non-terminal syntactic categories are given in angle brackets. Non-terminal **<&>** denotes a conjunctive word sequence, i.e. the following regular expression: **(as well as)|(and[/or])|(or[/and])**. Special characters **(**, **)**, **[**, **]**, **|**, and ***** have the usual interpretation in regular expression notation.

[6] In the experiments that we have performed, the precision of expanding terms from coordinated structures was 70%.

<TERM>([(]**such as**|**like**|(**e.g.**[,])) <TERM> (,<TERM>)* [[,] <&> <TERM>] [)]
<TERM> (,<TERM>)* [,] <&> **other** <TERM>
<TERM> [,] (**including**|**especially**) <TERM> (,<TERM>)* [[,] <&> <TERM>]
both <TERM> **and** <TERM>
either <TERM> **or** <TERM>
neither <TERM> **nor** <TERM>

Table 3: Sample of enumeration lexico-syntactic patterns

(<N>|<Adj>) (,(<N>|<Adj>))* [,] <&> (<N>|<Adj>) <TERM>
(<N>|<Adj>)/(<N>|<Adj>) <TERM>
(<N>|<Adj>) <TERM> (,<TERM>)* [,] <&> <TERM>
(<N>|<Adj>) <TERM>/<TERM>

Table 4: Sample of coordination patterns

head coordination	**[**adrenal **[**glands and gonads**]]**
term conjunction	**[**adrenal glands**]** and **[**gonads**]**

Table 5: Ambiguities of coordinated structures

2.4 Hybrid CLS Similarity

None of the similarities introduced so far is sufficient on its own to reliably estimate similarity between two arbitrary terms. For example, if a term appears infrequently or within very specific CPs, the number of its significant CPs will influence its contextual similarity to other terms. Further, there are concepts that have idiosyncratic names (e.g. a protein named `Bride of sevenless`), which thus cannot be classified relying exclusively on lexical similarity. Our experiments also show that syntactical similarity provides high precision, but low recall when used on its own, as not all terms appear in a parallel lexico-syntactical expression.

Therefore, we introduce a hybrid term similarity measure, called the *CLS* similarity, as a linear combination of the three similarity measures:

$$CLS(t_1, t_2) = \alpha\, CS(t_1, t_2) + \beta\, LS(t_1, t_2) + \gamma\, SS(t_1, t_2)$$

The choice of the weights α, β, and γ in the previous formula is not a trivial problem. In our preliminary experiments (Section 3) we used manually chosen values. However, the parameters have also been fine-tuned automatically by supervised learning method based on a genetic algorithm approach (Spasić et al. (2002)). A domain specific ontology has been used to evaluate the generated similarity measures and to set the direction of their convergence. The differences between results based on the various parameters are presented in the following section.

3 Results, Evaluation and Discussion

The *CLS* measure was tested on a corpus of 2008 abstracts retrieved from MEDLINE database (MEDLINE (2002)) with manually chosen values 0.3, 0.3 and 0.4 for α, β, and γ respectively. Random samples of results have been evaluated by a domain expert, and the combined measure proved to be a good indicator of semantic similarity. Table 6 shows the similarity of term `retinoic acid receptor` to a number of terms. The examples point out the importance of combining different types of term similarities. For instance, the low value of contextual similarity[7] for `retinoid X receptor` is balanced out by the other two similarity values, thus correctly indicating it as a term similar to term `retinoic acid receptor`. Similarly, the high value of the contextual similarity for `signal transduction pathway` is neutralised by the other two similarity

[7] The low value is caused by relatively low frequency of the term's occurrences in the corpus.

values, hence preventing it as being labelled as similar to `retinoic acid receptor`.

Term	*CS*	*SS*	*LS*	*CLS*
nuclear receptor	0.58	1.00	0.50	0.72
retinoid X receptor	0.32	1.00	0.33	0.60
retinoic acid	0.31	0.00	1.00	0.39
receptor complex	0.52	0.00	0.50	0.31
progesteron receptor	0.35	0.00	0.50	0.25
signal transduction pathway	0.75	0.00	0.00	0.22

Table 6: Similarity values between `retinoic acid receptor` and other terms

The combined measure also proved to be consistent in the sense that similar terms share the same "friends" (Maynard and Ananiadou (2000)). For example, the similarity values of two similar terms `glucocorticoid receptor` and `estrogen receptor` (the value of their similarity is 0.68) with respect to other terms are mainly approximate (Table 7).

Term	glucocotricoid receptor	estrogen receptor
steroid receptor	0.66	0.64
progesterone receptor	0.55	0.59
human estrogen	0.28	0.37
retinoid x receptor	0.27	0.36
nuclear receptor	0.30	0.33
receptor complex	0.31	0.33
retinoic acid receptor	0.27	0.28
retinoid nuclear	0.26	0.26

Table 7: Similarity values for `glucocorticoid receptor` and `estrogen receptor`

The supervised learning of parameters resulted in the values 0.13, 0.81 and 0.06 for α, β, and γ respectively (see Spasić et al. (2002)). The measure with these values showed a higher degree of stability relative to the ontology-based similarity measure. Note that the lexical similarity appears to be the most important and the syntactical similarity to be insignificant. The ontology used as a seed for learning term similarities contained well-structured, standardised and preferred terms which resulted in promoting the lexical similarity as the most significant. On the other hand, the *SS* similarity is corpus-dependent: the size of the corpus and the frequency with which the concurrent lexico-syntactic patterns are realised in it, affect the syntactical similarity. In the training corpus such patterns occurred infrequently relative to the number of terms, which indicates that a bigger corpus is needed in the training phase. In order to increase the number of concurrent patterns, we also aim at including additional patterns that describe appositions and implementing procedures for resolution of co-referring terms. We also plan to experiment with parametrising the values of syntactical similarity depending on the number and type of patterns in which two terms appear simultaneously.

The main purpose of discovering term similarities is to produce a similarity matrix to identify term clusters. In Nenadić et al. (2002b) we present some preliminary results on term clustering using the *CLS* hybrid term similarity measure. Two different methods (namely the nearest neighbour and the Ward's method) have been used, and both achieved around 70% precision in clustering semantically similar terms.

Conclusions and Further Research

In this paper we have presented a novel method for the automatic discovery of term similarities. The method is based on the combination of contextual, lexical and syntactical similarities between terms. Lexical similarity exposes the resemblance between the words that constitute terms, while syntactical similarity is based on mutual co-occurrence in parallel lexico-syntactic patterns. Contextual similarity is based on the automatic discovery of significant contexts through contextual pattern mining. Although the approach is domain independent and knowledge-poor, automatically collected patterns are domain specific and they identify significant contexts in which terms tend to appear. However, in order to learn domain-appropriate term similarity parameters, we need to customise the method by incorporating domain-specific knowledge. For example, we have used an ontology to represent such knowledge.

The preliminary results in the domain of molecular biology have shown that the measure

proves to be a good indicator of semantic similarity between terms. Furthermore, the similarity measure is consistent at assigning weights: similar terms tend to share the same "friends", i.e. there is a significant degree of overlapping between terms that are similar. These results are encouraging, as terms are grouped reliably according to their contextual, syntactical and lexical similarities.

Besides term clustering (presented in Nenadić et al. (2002b)), the similarity measure can be used for several term-oriented knowledge management tasks. Our future work will focus on the term classification and the consistent population and update of ontologies. However, specific term relationship identification that will direct placing terms in a hierarchy is needed. Further, term similarities can be used for term sense disambiguation as well, which is essential for resolving terminological confusion occurring in many domains.

Acknowledgement

We would like to thank Dr. Sylvie Albert and Dr. Dietrich Schuhmann from LION Bioscience for the evaluation of the results.

References

Ananiadou S. (1994): *A Methodology for Automatic Term Recognition*. Proceedings of COLING-94, Kyoto, Japan.

Bourigault D. (1992): *Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases*. Proceedings of 14th International Conference on Computational Linguistics, Nantes, France, pp. 977-981.

Frantzi K.T., Ananiadou S. and Mima H. (2000): *Automatic Recognition of Multi-Word Terms: the C-value/NC-value method*. International Journal on Digital Libraries, 3/2, pp. 115-130.

Friedman C., Kra P., Yu H., Krauthammer M. and Rzhetsky A. (2001): *GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles*. Bioinformatics, 17/1, pp. S74-S82.

Grefenstette G. (1994): *Exploration in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Massachusetts, p. 302.

Hatzivassiloglou V., Duboue P. and Rzetsky A. (2001): *Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach*. Bioinformatics, 17/1, pp. S97-S106

Hearst M.A. (1992): *Automatic acquisition of hyponyms from large text corpora*. Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France.

Klavans J. L., Tzoukermann E. and Jacquemin C. (1997): *A Natural Language Approach to Multi-Word Term Conflation.* Proceedings of Workshop DELOS, Zurich, pp. 33-40.

Maynard D. and Ananiadou S. (2000): *Identifying Terms by Their Family and Friends*. Proceedings of COLING 2000, Luxembourg, pp.530-536.

MEDLINE (2002): *National Library of Medicine*. http://www.ncbi.nlm.nih.gov/PubMed/

Mima H., Ananiadou S. and Nenadić G. (2001): ATRACT Workbench: *An Automatic Term Recognition and Clustering of Terms*. Text, Speech and Dialogue - TSD 2001, LNAI 2166, Springer-Verlag, Berlin, pp. 126-133.

Mima H., Ananiadou S., Nenadić G. and Tsujii J. (2002): *A Methodology for Terminology-based Knowledge Acquisition and Integration*. Proceedings of COLING 2002, Taiwan

Nakagawa H. and Mori, T. (2000): *Nested Collocation and Compound Noun for Term Recognition*. Proceedings of the First Workshop on Computational Terminology COMPUTERM 98, pp. 64-70.

Nenadić G., Spasić I. and Ananiadou S. (2002a): *Automatic Acronym Acquisition and Term Variation Management within Domain-specific Texts*. Proceedings of LREC 2002, Las Palmas, Spain, pp. 2155-2162.

Nenadić G., Spasić I. and Ananiadou S. (2002b): *Term Clustering using a Corpus-Based Similarity Measure*. Text, Speech and Dialogue - TSD 2002, LNAI series, Springer-Verlag, Berlin

Santini S. and Jain R. (1999): *Similarity Measures*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21/9, pp. 871-88

Spasić I., Nenadić G. and Ananiadou S. (2002): *Supervised Learning of Term Similarities*. IDEAL 2002, LNAI series, Springer-Verlag, Berlin

Ushioda A. (1996): *Hierarchical Clustering of Words*. Proceedings of COLING '96, Copenhagen, pp. 1159-1162.