Learning description of term patterns using glossary resources.

Le An Ha

School of Humanities, Languagues and Social Scienses
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB
United Kingdom
in6930@wlv.ac.uk

Abstract

In this paper, we describe a method of automatic extraction of knowledge patterns using in term descriptions from glossary, and using them to extract term description from technical texts. A brief introduction of the problem will be presented. After that, knowledge patterns and related works are discussed. By statistical analysis, we will show that these patterns can be learned, and we will propose a method to learn these pattern based on discover collocation of important verbs and nouns represent main concepts in the domain. Evaluation have been made showing that using the method, we can extract half of the correct descriptions, with a noise just one third.

1. Introduction

Together with the era of information, the number of technical texts is increasing exponentially. Now it is nearly impossible for terminologists to process these texts and produce a fair good, up-to-date terminology. There is a need for automatic terminology processing technique to be introduced, in order to keep up with that development. Recent advance in computational terminology offer some opportunities to solve the problem.

One of the main tasks that requires a lot of human labour is term description processing. In corpora based terminology processing, [finding, filtering, writing] description for terms is a very important and time-consuming task. A good term description should contain a information one user may need to know about the term, what it is, what are relations between it and other terms in the field, what makes it differ from other terms etc. We cannot completely name what information is needed; it much depends on domain, and also on the nature of users and applications. These descriptions can also contain information about the relationship between terms, which can be use for knowledge based term processing.

A term description can be considered as an extended version of term definition, with more information is added. That information is important in the field. For example, a description for ”chemical bond”, extracted from CHEM:

> A chemical bond is a strong attraction between two or more atoms. Bonds hold atoms in molecules and crystals together. There are many types of chemical bonds, but all involve electrons which are either shared or transferred between the bonded atoms.

The first sentence can be considered as a definition. It follows the definition formula: X = Y + distinguishing characteristic. But the following sentences give more information about the term, and seem not to follow any formula. It only depends on the domain, what is important when describing a term in that domain. The author of the glossary, who is an expert in the field, will decide which information to be included.

Recently, there are some proposed methods of (semi) automatically extracting description of term from real text, using knowledge patterns (Meyer, 2001). The approach will use these patterns to extract valuable information for terms. These patterns can be lexical patterns, contains certain words or words combination which are usually used in term description (for example: ”is a”, ”classified as”, ”form when”), they also can be grammatical patterns, or paralinguistic patterns. In this paper, we will concentrate in extracting domain-specific lexical patterns.

With this approach, there is a problem arising: how to acquire these knowledge patterns? Of courses, these knowledge patterns can be found by empirical observation of term descriptions, but it is a time-consuming task, and also domain-dependent. Different domains require different kind of information, thus have different knowledge patterns. For example, in chemistry, a description for a chemical compound may contain information about how it is formed, where it is used, what it contains, etc. But when we move into other domains, say weather, these questions about a term will change into how, when it occurs etc. The solution for the problem is using automatic techniques to learn them (knowledge patterns) from available resources which are rich in knowledge patterns, such as glossary, dictionary, encyclopedia and terminology . In these resources, the patterns are used extensively, thus easier to design an algorithm that can learn these patterns. In this paper, we will use glossaries as a training environment for the technique, because of their nature:

1) Less formal than other resources, thus containing more nature-occurring text.

2) Widely available in almost all domains.

Through statistical analysis, we will show that a glossary is a suitable resource to learn knowledge patterns, and suggest a method of extracting knowledge patterns from glossary, focusing on verbs expressing main relations and nouns representing main concepts in the domains.

2. Knowledge patterns and related works

2.1. Knowledge patterns

2.1.1. Some definitions and properties

Knowledge patterns(Meyer, 2001) (defining metalanguage (Pearson, 1998)) have different names designated by different researches depending on his/her own interest. For example, (Meyer, 2001) concentrates in conceptual relation and try to build a knowledge based terminology, defines them as certain predictable recurring patterns in text that conceptual relation will manifest itself in. While Pearson, more interested in patterns of definition, see them as a systematic way to fill in slots in formula of definition.

In this paper, we consider that knowledge patterns are patterns that people ”tend to use” when describe a term in a domain. It should has following properties:

1) repeated patterns.

2) appear in context that gives descriptions for term.

3) widely used in the domain.

This definition is close to Meyer definition, but, as we mention above, our target is the description of terms, not what it means, so we avoid mentioning about conceptual relations.

2.1.2. Types of knowledge patterns

Knowledge patterns can be classified into different types by their linguistic properties. They can be lexical, grammatical, or paralinguistic patterns. More about types of knowledge patterns can be found in the Meyer’s work.

2.2. Related work

(Fujii and Ishikawa, 2000) suggest using Encyclopedia to extract these patterns in Japanese They extract the co-occurrence of bunsetsu phrases and post-edited the extracted patterns. They use these patterns, combining with tri-gram language model to extract term descriptions from the World Wide Web. They also benefit from html tags such as DD and DT, which are inherently provided to describe terms. Their method is domain independent (in fact, they try to avoid domain-specific patterns). Interestingly, we find out that our results are similar to theirs, although our method is for English, and different from what they suggest.

(Morin, 1999) use the (Hearst, 1998) frame-work to extract patterns for synonyms and hypernyms relations. This technique is quite reliable but based on resources of known relationship between terms (semantic network). As we can see that a semantic network is not always available in a certain domain, also there are a lot of relations which can not be generalized into some known relations, such as HYPONYM and MERONYM, but very important in a certain domain. The frame-work also focus in highly unambiguous patterns, ”NP0 such as NP1, NP2....”, which may not be domain-specific. Morin calculates similarities between two lexical-syntactic expressions by longest common string, and uses them for clustering patterns, and then uses the clustered patterns to extract semantics relations between terms from technical corpora.

(Pearson, 1998) conducts an empirical observation of term definitions in context, base on definition formula: X = Y + distinguishing characteristic, and tries to specify slot fillers for X, Y, = in that formula by analysing the three corpora. For Y, she suggests the class words such as ”technique”, ”method”, ”process”, ”function”, ”property”. For =, the main focus is in the use of connective verb, such as ”comprise”, ”consist”, ”define”. The analysis also shows interested points about the use of focusing adverbs. In that research, Pearson tries to discover domain-independent patterns of definition.

Works (Boguraev et al., 1989) also have been done in analysing dictionary definition, show that there are certain patterns which have been used for definition.

3. Glossaries and their statistical properties

3.1. Definition and some properties

A glossary is ”an alphabetical list of special, usual or technical words or expressions, giving their meanings”. (Collins Cobuild English Dictionary 1998)

A glossary is usually compiled by an expert in the field. In the past, it is usually published as an appendix of a technical text, thus may be brief and considered personal and informal. But today, the situation has changed. When looking for information on the internet, we can notice that website of a specific domain usually has a glossary, and people also take advance of hyper text features, using hyper link between terms, and multimedia explanation for terms. This change makes glossaries become inexpensive, but valuable resources for natural language processing.

A glossary should have the following properties:

1) An alphabetical list of technical terms in a specialized field of knowledge.

2) Giving meanings, definitions of the items and other information which are widely used in the domain.

3.2. Statistical analysis

We have collected different glossaries from different domain from the world wide web, and use FDG parser (Tapanainen and Jarvinen, 1997) to tag them, the FDG parser also give some information about grammatical functions, which can be used in further analysis, but in this paper, we only use the part-of-speech tagging information. Summary of glossaries being analysed can found in table 1.

We examine the frequency list of verbs, nouns, adjectives to see the different between glossaries and other kinds of texts. Using BNC as a reference, we notice the following:

The use of verbs: examine the frequency list of verbs using in glossaries, we notice that: there is a different between glossaries and other kinds of text in the use of verbs, the use of verbs in glossary reflects main relations between terms in the domains. Also in different glossaries, the uses of verbs are also different. For example, in chemistry, when observe the verb frequency list, we can see that ”use”, ”contain”, ”form”, etc. are important relations, and ”use”, ”occur”, ”form”, ”refer”, ”include” etc. are widely used in descriptions of terms in weather. Note that, ”refer” and ”include” may be domain-independent.

In the beginning, we expect that the ”connective verbs” as described in (Sager, 1990) and (Pearson, 1998) should appear in the high frequency list of verbs, but the results show that they do not. The reason is the glossary’s editor starts the descriptions directly, instead of using these verbs. The evaluation also shows that these above ”relations verbs” are as important as ”connective verbs” in description of terms.

Glossary	Sources	Domain	No of terms	No of words
AGRI	www.cnie.org	agriculture	1100	59393
CHEM	chemed.chem.purdue.edu	chemistry	1042	38733
WEATHER	www.weather.com	weather	2145	36898

Table 1: Summary of glossaries in analysis

verb	frequency	proportion	BNC proportion
be	744	0.23	0.22
use	118	0.03	0.00066
have	106	0.03	0.07
contain	75	0.02	0.0009
form	61	0.01	0.0009
make	55	0.01	0.01
produce	50	0.01	0.001
call	40	0.01	0.002
bind	33	0.01	0.0003
dissolve	32	0.01	0.00008
absorb	31	0.009	0.0001
involve	28	0.008	0.001
measure	27	0.008	0.0003
occur	27	0.008	0.00008
add	25	0.007	0.0001
react	24	0.007	0.0001
change	24	0.007	0.001

Table 2: The high frequency verbs list extracted from CHEM and their proportion, compare to their proportion in BNC

verb	frequency	proportion	BNC proportion
occur	69	0.02	0.0008
form	58	0.02	0.0009
measure	33	0.01	0.0003
develop	29	0.01	0.001
move	28	0.01	0.003
characterize	23	0.008	0.00005
fall	22	0.008	0.001
determine	20	0.0007	0.00006
blow	19	0.0007	0.00003
rise	17	0.0006	0.00007

Table 3: The high frequency verbs list extracted from WEATHER and their proportion, compare to their proportion in BNC

be *	use	provide	authorize
make *	include	require	establish
have *	receive	refer	pay *
sell	administer	apply	call

Table 4: The high frequency verbs list extracted from AGRV (* : also have high frequency in BNC, not statistically significant)

The analysis of the use of nouns is lead to a raw classification of concepts in the field, such as chemical compound {base, salt, polymer, hydroxide, acid etc.}, state of matter {solid, liquid, gas}, physical phenomenon {force, pressure, entropy, field, radiation} etc. A more careful analysis of the nouns at the beginning of the description may lead to a list of domain-specific class words, which are the generic term in the domain, such as ”reaction”, ”acid”, ”substance”, ”compound” in chemistry, and ”air”, ”wind”, ”pressure”, ”temperature” in weather. When comparing to the list of class words suggested by Pearson, which are domain-independent, we can see that more domain-specific class words can be automatically extracted from glossary, for using in term description extraction.

AGRI	CHEM	WEATHER
program	molecule	air
act	ion	wind
price	atom	term
commodity	reaction	pressure
food	acid	cloud
farm	substance	temperature
state	solution	surface
land	compound	water
water	water	earth
service	electron	weather
payment	energy	area
trade	gas	atmosphere
crop	bond	ice
product	temperature	snow
market	pressure	wave

Table 5: The first 15 nouns extracted from frequency lists of the three glossaries in analysis

When examining the adjective frequency list of these glossaries, we can see that some of the domain-specific adjective have been intensive used, together with other the general adjectives, such as chemical, atomic, molecular in chemistry and in weather, they are high, low, cold, warm [1] and atmospheric.

One can note from these above tables that AGRI is somehow strange, and the reason is that this is ”Glossary of Terms, Programs, and Laws”, so it does not only contain terms from agriculture, but also state programs and laws. One has to be careful when choosing a suitable glossary for their own work.

[1] these adjective themselves are not domain-specific, but can be used combining with other indicator

AGRI	CHEM	WEATHER
national	chemical	related
agricultural	atomic	low
federal	molecular	high
rural	different	cold
united	constant	atmospheric
low	equal	warm
natural	strong	tropical
eligible	same	opposite
domestic	solid	small
certain	standard	great

Table 6: The first 10 adjectives extracted from frequency lists of the three glossaries in analysis

4. Learning knowledge patterns

Learning lexico-syntactic patterns always a problem which was addressed in previous works of (Hearst, 1998; Morin, 1999; Manning, 1993). We had originally planned to use Morin method to learn lexico-syntactic patterns, but when implementing the method, we notice following problem:

1) The lack of data: with each verb, there are only quite a few examples to use for training.

2) The lack of similarity measures: the measures described by Morin are not the best suitable one for lexico-syntactic patterns, and a good measure is still to be introduced.

Then we come back to a more simple idea, considering the verb itself as a knowledge pattern, and use statistics to decide which verbs are included in the knowledge pattern list. We use relative frequency ratio (Damerau, 1993), with BNC as a reference corpus; we set a threshold of 10 from empirical observation. All verbs that have relative frequency ratio bigger than 10 should be considered as knowledge pattern, other uncertain verbs (relative frequency ratio around 10), we will use context of these verbs, which is extracted from glossary, as addition criteria, for example, with the verb ”produce” in CHEM, relative frequency ratio is 10, we extracted the left context of ”produce” from CHEM, which are ”gas”, ”solution”, ”substance” etc., and right context, which are ”reaction”, ”water”, ”ethanol”, ”combustion”, ”solution” etc. We will try to cluster these nouns into groups having the same hypernym (water, ethanol are chemical compound) using Wordnet. Knowledge patterns formed from these verbs will include their context.

5. Results and evaluation

Unlike other research topics, in the field of automatic terminology processing, there is no ”gold standard” for evaluation, and large scale evaluation may be expensive. We then choose a sample set of 14 terms from chemistry and 8 from weather domain, and then extract randomly a set of sentences contain these terms from World Wide Web, the system will decide a sentence contains description of the term if it contains (a) knowledge pattern(s) extracted from the above process. Human experts will judge the decision of the system.

For example: these following sentences contain ”nucleotide base”. In which, (2), (3), (4), (8) are descriptions of ”nucleotide base”, and the system have chosen (1), (2), (3), (4), (8) as results, because of ”found in”, ”occur”, ”form”, ”contain”, ”form” are considered as knowledge patterns learned from CHEM.

(1) If you are familiar with chemical diagrams, the image below shows the four nucleotide bases found in DNA:

(2) While there are only 4 different nucleotide bases that can occur in a nucleic acid, each nucleic acid contains millions of bases bonded to it.

(3) There are four different nucleotide bases that occur in DNA: adenine (A), cytosine (C), guanine (G) and thymine (T).

(4) The nucleotide bases of the DNA molecule form complementary pairs: the nucleotides hydrogen bond to another nucleotide base in a strand of DNA opposite to the original.

(5) While DNA cloning into a plasmid allows the insertion of DNA fragment of about 10,000 nucleotide base pairs, DNA cloning into a YAC allows the insertion of DNA fragments up to 1,000,000 nucleotide base pairs.

(6) Help students see that this process is similar to that of using modified nucleotide bases for DNA sequencing.

(7) Each strand of DNA contains millions or even billions (in the case of human DNA) of nucleotide bases.

(8) Each nucleotide base in the DNA strand will cross-link (via hydrogen bonds) with a nucleotide base in a second strand of DNA forming a structure that resembles a ladder.

(9) The order of nucleotide bases in a DNA molecule; determines structure of proteins encoded by that DNA.

terms	#S	#D	#E	#C
ionic bond	19	7	6	5
acetid acid	12	3	1	0
acid-base indicator	10	4	1	1
Charles’s Law	10	4	4	4
Diazonium salt	10	2	6	2
eutectic mixture	8	1	1	1
hygroscopicity	13	0	0	0
isomerization	10	1	2	0
nucleotide bases	9	4	5	4
photosynthesis	10	6	3	3
polar molecule	12	3	2	1
position of equilibrium	3	1	0	0
qualitative analysis	10	2	2	1
unimolecular reaction	10	2	2	1
Total	146	40	35	23

Table 7: Results of 14 chemistry terms (#S: number of sentences in analysis, #D: number of sentences containing description for term, #E: number of sentences extracted as term descriptions by the system, #C: number of correct descriptions)

The results show that the proposed method can extract sentences contain term description with practical rate of success, two thirds of the extracted descriptions are correct, also it can discover half of the descriptions.

6. Conclusion and future works

In this paper, we show that, by statistically analysing a glossary, we can extracted valuable domain-specific linguistic knowledge. We use statistic method to extract

terms	#S	#D	#E	#C
adeabatic process	13	7	4	3
anomalous propagation	10	2	1	1
astronomical light	11	5	3	3
black blizzard	6	1	1	0
constanst pressure surface	8	3	3	2
equatorial trough	12	4	4	3
dew point	11	6	3	2
indian summer	12	2	2	2
total	83	30	21	14

Table 8: Results of 8 weather terms (#S: number of sentences in analysis, #D: number of sentences containing description for term, #E: number of sentences extracted as term description by the system, #C: number of correct description)

knowledge patterns from a glossary, and use them to extract description of terms from technical texts. The results show that the proposed method has a practical use, in average, two thirds of the extracted description are correct, cover half of the descriptions appear in the test set.

The future work will include extensively evaluating, exploring other indicators for knowledge patterns to improve the performance.

7. References

B. Boguraev and T. Briscoe. Ed. 1989. *Computational lexicography for Natural Language Processing.* London. Long Man.

F. J. Damerau. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29:433–447.

M. A. Hearst 1998. Automatic discovery of wordnet relations. In C. Fellbaum editor, *Wordnet: An electronic Lexical database*, 131-151. Cambridge, M.A. MIT press.

A. Fujii and T. Ishikawa. 2000. Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts. *ACL 2000*, 488–495.

C. D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. *ACL 93*, 235–242.

I. Meyer. 2001. Extracting knowledge-rich contexts for terminography In D. Bourigault, C. Jacquemin and M. C L'Homme editors, *Recent Advances in Computational Terminology*, Amsterdam,John Benjamins.

E. Morin 1999. Automatic acquisition of semantic relations between terms from technical corpora. *TKE99*, 268–278.

J. Pearson. 1999. *Terms in context.* Studies in Corpus Linguistics. Amsterdam: John Benjamins.

J. C. Sager. 1990. *A practical course in terminology processing.* Amsterdam. John Benjamins.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

P. Tapanainen and T. Jarvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applies Natural Language Processing*, 64–71.