自然言語処理 112－12
（1996. 3. 15）

専門用語の自動抽出

Katerina T. FRANTZI†, Sophia ANANIADOU†, 辻井潤一‡

†NTT コミュニケーション科学研究所
{katerina,sophia}@nttkb.ntt.jp

‡東京大学理学部情報科学科
tsujii@is.s.u-tokyo.ac.jp

本論文では、医学コーパスから医学における専門用語を自動的に抽出する技法について議論する。最終的な手法においては、言語学的な知識と統計的知識の融合を目指すが、本論文では、主として、統計的な手法について報告する。特に、複数語からなる表現の専門用語らしさを評定する尺度を新たに提案し、その有効性を RECALL 率、precision 率の 2 つの基準で評価した結果について、報告する。

Extracting Terminological Expressions

Katerina T. FRANTZI†, Sophia ANANIADOU†, Junichi TSUJII‡

†NTT Communication Science Laboratories[1],
{katerina,sophia}@nttkb.ntt.jp

‡University of Tokyo,
Dept. of Information Science
tsujii@is.s.u-tokyo.ac.jp

In this paper we deal with the problem of automatic term recognition in medical corpora combining linguistic knowledge with statistics. We propose a statistical measure which extracts multiword, uninterrupted terms. It also deals with those terms being substrings of other longer terms. We evaluate the results using precision and recall.

[1]Permanent address: Manchester Metropolitan University, {K.Frantzi,S.Ananiadou}@doc.mmu.ac.uk

1 Introduction

Term recognition is of great importance as the existence of coherently built terminologies leads to better performance for many interesting applications. Human-based translation, machine translation, text categorization and indexing in digital libraries, technical writing, book indexing, information retrieval and hypertext linking and authoring to name a few. However, it is a time consuming task due to the many problems involved.

Approaches for term recognition are mainly based on linguistic knowledge (Bourigault, 1992; Ananiadou, 1994). Linguistic rules (ie. in the form of morphosyntactic properties) are applied to sublanguage corpora.

The availability of large scale corpora enhanced the use of statistical approaches on natural language processing in general, for inferring grammars or enriching lexica. The importance of corpora cannot be denied, since the pertinent information is obtained by the *real data*. Most term recognition approaches relied on the criterion of frequency of occurrence (Dagan and Church, 1994; Daille et al., 1994; Justeson and Katz, 1995). In this paper we are concerned with the extraction of multi-word, uninterrupted terms. We argue that the frequency of occurrence as the only statistical measure is not enough and propose an alternative.

2 Term Recognition

A term is the "designation of a defined concept in a special language by a linguistic expression" (ISO, 1988). Terms are strongly connected to the sublanguages they occur, each of which has its own set of concepts and is related to a particular subject field. Sublanguages demonstrate restrictions on the lexical, syntactic and semantic level. However, a general language grammar and lexicon are not enough for their description since sublanguages are not just subsets of the general language. We can sup up the importance of automatic term recognition as follows:

1. the variety of applications (mentioned in introduction).
2. the fact that new terms are constantly created.

The terms, are ideally monoreferential: a concept should be descripted by one term only, and a term should describe only one concept. However, this is not always the case. Polysemy, homonymy and synonyny make term recognition a difficult task (Lauriston, 1994). As an example, consider the following:

Head of the human body (medicine)
Head of a tool (engineering
Head of a plant (botanology)
Head of a noun phrase (linguistics)

The availability of large scale corpora influenced term recognition approaches, which can be divided into 3 categories:

1. Linguistic approaches (i.e. using morphosyntactic criteria)
2. statistical approaches (using statistical information from the corpora)
3. hybrid approaches (combining both linguistic and statistical knowledge)

3 A Hybrid Approach

We have adopted a hybrid approach, combining linguistic knowledge and statistics since we cannot deny the importance of both. Our domain of application is eye pathology patient reports. We derived after examining the term formation patterns

the following simple grammar which acted as a linguistic filter:

(ADJECTIVE | NOUN)*NOUN

This grammar should be eventually validated by the results of our technique. A stop list was progressively build by examining the resulting lists of our measure. This stop list acted as an additional filter over the candidate terms.
Our statistical measure uses 3 factors:

1. the n-gram's total frequency of occurrence in the corpus,
2. its frequency of occurrence in longer (already extracted) candidate terms,
3. the number of these longer candidate terms.

The importance of the first factor is that technical terms tend to appear with relatively high frequency of occurrence. We must keep in mind though, that only frequency of occurrence is no guarantee of termhood. The second and third factors exist in order to prevent substrings of terms to be erroneously extracted as terms due to their high frequency of occurrence. Consider the following:
soft contact lenses
hard contact lenses
contact lenses
soft contact
If we had relied simply on the frequency of occurrence, then we would have extracted not only *soft contact lenses* and *hard contact lenses* as candidate terms but also *contact lenses* and *soft contact*, since both of the latter have at least the same frequency as their supersets. In addition they are both tagged as Noun–Noun.

The problem is that while *contact lenses* should be extracted as a candidate term, *soft contact* should not. When a substring is a candidate term on its own? What is important is the degree of independence of the examined n-gram. *Contact lenses* can occur in isolation, while this is not the case for *soft contact*. If it does not appear in longer candidate terms, then it is more probable of being a term on its own. If it appears in longer, already extracted, terms, then the higher the number of the longer candidate terms within it occurs is, the more probable it is for the substring to be a candidate term on its own.

We use a measure, which is the combination of the three above mentioned factors. We call this measure *C-value*.

$$C\text{-}value(a) = n(a) - \frac{t(a)}{c(a)} \qquad (1)$$

where
a is the examined n-gram,
$n(a)$ is the total frequency of occurrence of a on the corpus,
$t(a)$ is the frequency of occurrence of a in longer (already extracted) candidate terms.
$c(a)$ is the number of those candidate terms.
In the case where a is not part of a longer candidate term, then it is assigned

$$C\text{-}value(a) = n(a) \qquad (2)$$

It is straightforward that if a has the same frequency with one longer candidate term that contains a, it is assigned $C\text{-}value(a) = 0$ i.e. is not a term. We must keep in mind though that the above factors describe probabilities. That is, the measure attempts to extract the most-probable-to-be-terms n-grams, keeping both precision and recall in relatively high values.

4 The Experiments

For our experiments we used a corpus consisting of medical records on eye pathology. The size of the corpus is 860,000 words. We tagged the corpus using Brill's part-of-speech-tagger (Brill,

C-value	String	*C-value*	String
2116.55	OPTIC/JJ NERVE/NN	594.345	CORNEAL/JJ DISC/NNP
1404.45	FIBROUS/JJ TISSUE/NN	530.806	BASAL/NNP CELL/NN
1201	ANTERIOR/JJ CHAMBER/NN	525.682	NERVE/NN HEAD/NN
1129.67	CORNEAL/JJ DIAMETERS/NNS	523.88	BASAL/NNP CELL/NN CARCINOMA/NN
1013.55	CELL/NN CARCINOMA/NN	491.861	CELL/NNP PAPILLOMA/NNP
974.85	LEFT/NNP GLOBE/NNP *	487.16	LEFT/NNP UPPER/NNP *
897.6	TEMPORAL/JJ SIDE/NN	460	BASAL/JJ CELL/NN CARCINOMA/NN
890.818	PLASMA/NN CELLS/NNS	440.607	LEFT/NNP UPPER/NNP LID/NN
863.755	UPPER/NNP LID/NN	439.875	BASAL/JJ CELL/NN
743.1	MALIGNANT/JJ MELANOMA/NNP	415	HYALINE/JJ FIBROUS/JJ TISSUE/NN
665	NASAL/JJ SIDE/NN	409.25	AXIAL/JJ REGION/NNP
636.723	LEFT/NNP EYE/NNP *	400.455	SUBSTANTIA/NNP PROPRIA/NNP
628.3	TRABECULAR/NNP MESHWORK/NNP	366.118	BASAL/JJ CELL/NNP
621.84	LID/NN MARGIN/NN	360.5	WHITE/NNP NODULE/NNP
594.508	RETINAL/JJ DETACHMENT/NNP	354.917	WHITE/NNP EXCRESCENCE/NNP

Table 1: First 30 n-grams extracted with *C-value*. * stands for non–term.

Threshold	Acc. Precision	Recall	C-value Intervals	Interval–Precision
53	89.5%	76%	2197–53	89.5%
20	76.7%	80%	50–20	68.2%
13	72.9%	82%	20–13	63.4%
10	69.7%	83.4%	13–10	60.0%
5	62.0%	89.2%	10–5	48.5%
1	34.2%	90.5%	5–1	33.4%

Table 2: Precision and Recall

1992). We have not retrained the tagger yet. This will be required as our corpus is in capital letters and there are problems caused by that. From Table 1 one can see that "BASAL CELL" appears three times, as "BASAL/NNP CELL/NN", as "BASAL/JJ CELL/NN" and as "BASAL/JJ CELL/NNP" spliting the *C-value* of "BASAL CELL", while it should really appear once.

We first extracted the n-grams for n=2,3,4,5. The length of the n-grams though is variable. The output n-grams are filtered such that only those that are satisfied by the above grammar (and the stop list) are passed to the second step which is the calculation of the *C-value* for each of these n-grams. The values for the longer n-grams are calculated first. The extracted n-grams, which are the candidate terms, are sorted according to their *C-value*. Table 1 shows the first 30 n-grams extracted. The results were evaluated by a terminologist and domain–expert working on samples of the corpus and on samples of the extracted list of candidate terms. We used Precision and Recall for the evaluation of the results. Precision is defined at the ratio of the retrieved valid elements over the total number of the retrieved elements. Recall, as the ratio of the retrieved valid elements over the total number of valid elements. These two measures were calculated for various thresholds of *C-value*. For the precision both the accumulated value as well as that on *C-value* intervals were estimated. Since both precision and recall could not be calculated on the whole corpus due to its size, we estimated these values using samples. Table 2 gives a sum-

mary. Column 1, is the threshold for *C-value*. Column 2 is the Accumulated Precision, and Column 3 the Recall for these thresholds, while Column 5 is the Precision for *C-value* being within the intervals shown in Column 4. The place of the threshold point depends on the user's application. It is straightforward that a threshold set on high values increases the precision but decreases the recall and vice versa.

5 Related Work

Linguistic approaches to term recognition are reported on work by (Bourigault, 1992; Ananiadou, 1994; Spyns, 1994). Bourigault's LEXTER extracts noun phrases which due to their grammatical structures and their place in maximal-length noun phrases are likely to be terminological units. Ananiadou implemented a computational term formation grammar using a morphological analyser which recognises monomorphemic, derived and compound terms from the medical field of Immunology. Spyns implemented a large scale category guesser for Dutch medical language (cardiology), making use of non-morphological (endstring matching) as well as morphological knowledge. The above mentioned methods, lack of any statistical information gained from the real data.

Three recent hybrid approaches are those by (Daille et al., 1994; Justeson and Katz, 1995; Dagan and Church, 1995). Daille et al., use linguistic knowledge to identify certain noun phrases, in English and in French, which are likely to be terms. Then, they apply different statistical scores to the candidate terms to extract the terms. The statistical measure they end up with is the frequency of occurrence. The corpus consists of 200,000 words of each language in the area of telecommunications, and the terms extracted have length 2 or more. Dagan and Church extract sequences of nouns that do not appear in a standard stop-list. The frequency of occurrence produces the lists of candidate terms; these lists are then checked by terminologists. This approach seems to have high recall since only a stop list keeps words from appearing as candidate terms but it lacks in precision for the same reason. Finally, Justeson and Katz present a terminology identification algorithm that is motivated by linguistic properties of terms. The candidate strings have a frequency of 2 or more and are described by a regular expression. Their method is domain-independent.

6 Conclusions

In this paper we presented an approach to extract multi-word, uninterrupted terms from the medical field of eye pathology. Our approach combines both linguistic and statistical knowledge gained from the *real data*. The linguistic knowledge is given on the form of a simple grammar. The statistical measure though based on the frequency of occurrence, deals with the problem of substrings of terms.

In future we are planning to tune the grammar and stop list and add term formation morphological rules. We also plan to change the parameters of the statistical measure and evaluate the results. Therefore, besides being more precise on the estimation of precision and recall (using bigger samples), we also expect to get better results. Finally, we are going to apply the method to some other field (i.e. market reports).

7 Acknowledgements

We would like to thank Dr. Tom Sharpe from the University of Manchester for providing us with the corpus.

References

Ananiadou, S. 1994. A Methodology for Automatic Term Recognition. In *Proceedings of COLING*, pages 1034–1038.

Bourigault, D. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of COLING*, pages 977–981.

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proc. of the Third Conference of Applied Natural Language Processing, ACL*, pages 152–155.

Dagan, I. and Church, K. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of EACL*, pages 34–40.

Daille, B.; Gaussier, E. and Lange, J.M. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of COLING 94*, 515-521.

International Organization for Standardization (ISO). 1988. Terminology - vocabulary. In *(ISO/DIS 1087)*.

Justeson, J.S. and Katz, S.M. 1995. Technical terminology:some linguistic properties and an algorithm for identification in text. In *Natural Language Engineering*, 1:9–27.

Lauriston, A. 1994. Automatic recognition of complex terms: Problems and the TERMINO solution. In *Terminology*, 1:147–170.