Clustering of Medical Terms based on Morpho-syntactic Features

Agnieszka Mykowiecka and Magorzata Marciniak
Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland

Keywords: Terminology Extraction, Term Clustering, Medical Data, Ontology.

Abstract: The paper presents the first results of clustering terms extracted from hospital discharge documents written in Polish. The aim of the task is to prepare data for an ontology reflecting the domain of documents. To begin, the characteristic of the language of texts, which differs significantly from general Polish, is given. Then, we describe the method of term extraction. In the process of finding related terms, we use lexical and syntactical information. We define term similarity based on: term contexts; coordinated sequences of terms; words that are parts of terms, e.g. their heads and modifiers. Then we performed several experiments with hierarchical clustering of the 300 most frequent terms. Finally, we describe the results and present an evaluation that compares the results with manually obtained groups.

1 INTRODUCTION

Term clustering is an indispensable component of the analysis and exploration of data, such as labeling data with semantic information, or the development of a domain ontology. Most of publications concerning this problem describe methods using big text corpora, e.g. (Ushioda, 1996), (Lin and Pantel, 2001). A review is given in (Cimiano, 2006). But there are also papers where data are relatively small, like (Le Moigno et al., 2002) which describes terminology aggregation from surgical intensive care documents, or an experiment where pulmonology data were analysed (Baneyx et al., 2006). In this latter case not only clinical data were taken into account but also a corpus created from a pulmonology book. Our experiment also concerns small and specific data.

Research done in the field of unsupervised terminology extraction assumes that similar terms share similar linguistic contexts so the extracted terms can be grouped using the morphosyntactic features of their neighboring words. In some attempts the additional semantic knowledge is also utilized—mostly in a form of Wordnet relations, e.g. (Navigli et al., 2003), (Ittoo and Maruster, 2009) or domain specific ontologies like SNOMED-CT (Pedersen et al., 2007). In case of the chosen specialized domain and multiword terminology, the coverage of Polish Wordnet is not high, so we decided to use only morphosyntactic information.

In the paper we present the first results of the clustering of medical terminology extracted from hospital discharge documents written in Polish. These documents include terms related to several different topics, i.e. anatomy, medicine, and health care. The aim of the task is to prepare data for establishing an ontology suited to this multi-topic domain. For now, we limit ourselves to recognizing groups of concepts, but not relationships between them. We begin our task with the automatic recognition of domain terms from a corpus using a shallow syntactic grammar. Then we determine term similarities on the basis of their lexical and syntactic features as well as contexts in which they occurred following the ideas presented in (Nenadić et al., 2004). From the top part of the list of ranked candidates of terms, we have selected 300 terms which took part in a clustering experiment. We excluded candidates that are not correct terms and those which are infrequent (below 8).To evaluate the results, the terms were independently clustered manually by 2 persons.

2 LINGUISTIC ANALYSIS

In our experiment we process medical data containing hospital discharge documents. They were collected from a surgical ward of a children's hospital. The set consists of 1165 documents, and contains over 380,000 tokens. The following aspects make task realization difficult: the data are not big, texts are noisy, and vocabulary is very specific. On the other hand,

Mykowiecka A. and Marciniak M..
Clustering of Medical Terms based on Morpho-syntactic Features.
DOI: 10.5220/0004137502140219
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 214-219
ISBN: 978-989-8565-30-3

physicians usually use expressions which are to some degree fixed, so the same patterns repeat in texts frequently, which probably makes our task tractable.

The vocabulary of clinical documents significantly differs from general Polish texts. The documents include many proper names e.g. names of medication *Furagin* or *Detromycyna*. There are also many acronyms and abbreviations, e.g. *TK* 'CT' (Computed Tomography). Some abbreviations are created ad hoc, e.g. *j. brzusznej* ('abdominal cavity') where *j.* here is the abbreviation of 'cavity', and can be properly recognized only in context. Diagnoses and bacterial names are frequently written in Latin.

The discharge records are not meant to be published, thus they are not carefully edited. The majority of errors are in words that are not included in the standard editor dictionary, like *elaktroresekcji* instead of *elektroresekcji* 'electroresection$_{gen}$' but in common words spelling errors are also quite frequent. A typical error is the lack of Polish diacritics, e.g. *miesiace*$_{nom}$ instead of *miesice*.

To recognize/identify candidates for ontology concepts from texts, an initial linguistic analysis of the texts is performed. It consists of:

- Segmentation into tokens. We distinguish words, numbers and punctuation marks.
- Morphological annotation. To each word we assign: its base form, part of speech, and complete morphological characterization. The annotation is based on the results obtained by the tagger TaKIPI (Piasecki, 2007) that utilizes the morphological analyzer Morfeusz SIAT (Woliski, 2006) and the *Guesser* module which suggests tags for words which are not in the dictionary.
- Correction of the annotation. We manually prepared a set of (context-free) correction rules applied to the already tagged data.
- Removing improperly recognized sentence endings after abbreviations, and adding end-of-sentence tags at the ends of paragraphs.

A detailed tagset comprises hundreds of tags describing about thirty grammatical classes and several morphological features. For total 11363 token types, there were 3676 different nominal forms while only 113 verbal forms were observed.

3 TERMS EXTRACTION

The aim of the presented research was to recognize how to group terms which occur in the selected type of text. The first step needed to achieve this goal is to identify the terms themselves. The decisions made at this stage are crucial for the results of the next processing steps. What should be considered as a term heavily depends on an adopted definition of the domain and the accepted degree of specificity. We propose a method of extracting terms directly from the documents, basing on their morphological features.

What is common to all domain vocabularies is that the vast majority (if not all) of the terms are noun phrases. Their internal structure can vary, but the types of constructions are limited. In Polish, domain terms most frequently have one of the following syntactic structures:

- a single noun or an acronym, e.g. *nerka* 'kidney', *USG*;
- a noun followed (or, more rarely, preceded) by an adjective, e.g. *czerwone*$_{adj}$ *krwinki*$_{n}$ 'red cells' *prawa*$_{adj}$ *nerka*$_{n}$ 'right kidney';
- a noun followed by another noun in genitive, e.g. *zapalenie*$_{n,nom}$ *puc*$_{n,gen}$ 'pneumonia';
- a combination of the last two structures, e.g. *zamanie*$_{n,nom}$ *prawej*$_{adj,gen}$ *rki*$_{n,gen}$ 'right hand fracture'.

The rules become more complicated if we want to take into account additional features of Polish nominal phrases: word order, genitive phrase nesting, prepositional modifiers and coordination. However, more complicated constructions usually do not describe one concept but a relation between two or more concepts. Thus, during the first phase of model creation we analyse only simple noun phrases. For recognizing the selected types of nominal phrases, a cascade of three simple shallow grammars was created.

Applying the adopted set of rules to the data, resulted in obtaining 4485 types of phrases (3404 top level types) which occurred 89839 times. For the resulting set of phrases, we performed an analysis similar to that proposed in (Frantzi et al., 2000) to identify subphrases which constitute separate terms (e.g. the phrase *pchrzyk ciowy prawidowy* 'normal gall bladder' describes a physician's judgment on a particular body part) and to rank the terms according to their importance measured in terms of usage frequency. We used a slightly modified definition of C-value given below (p – is a phrase under consideration, LP – is a set of phrases containing p, and ||LP|| – the number of phrases differing in elements which are adjacent to p):

$$C(p) = \begin{cases} lc(p) * freq(p) - \frac{1}{\|LP\|} \sum freq(lp) \\ \quad if\ \|LP\| > 0,\ lp \in LP \\ lc(p) * freq(p), if\ \|LP\| = 0 \end{cases}$$

where lc(p) = log$_2$(length(p))
if length(p) >1 and 0.1 otherwise.

As a result of the above procedure we obtained about 1500 terms for which the C-value was greater than 1. However, many of these terms occurred a small number of times in one or at most two different contexts. To limit the influence of singular occurrences of some terms and to make it possible for humans to evaluate the results in a reasonable time, we limited our experiment to the 300 terms taken from this ranked list. 15 elements form the top section of the list were replaced by subsequent elements from the list. We excluded several ambiguous or difficult to cluster acronyms like sex codes (M,K), 3 terms judged to be poor domain terms e.g.: *oglny mocz* ‘general urine’ (a part of the phrase in genitive *badania oglnego moczu* ‘general urine test’) or those terms which appear alone (not as a subphrase) less than 8 times, e.g.: *wskanik* ‘index’ that appears 5 times alone and 763 times in *wskanik protrombinowy* ‘prothrombin index’. To automatically group these selected elements, an adequate similarity measure was defined.

4 TERM SIMILARITY

4.1 Context

One type of similarity proposed in (Nenadić et al., 2004) is called contextual similarity, and is based on the contexts in which terms appear. For our purpose we consider left and right contexts of terms separately. Contexts are not allowed to cross sentence or paragraph boundaries. We decided to take into account the following types of context patterns:

- The form of the nearest left neighboring word.
- POS contexts. In this case patterns are strings of part of speech tags. We took into account patterns of 2 to 4 elements. If sentence boundaries are encountered, the context is shorter.
- The base form of the token preceding and following the term (separately).
- The base form of the nearest verb. If there are no verbs encountered within the sentence boundaries, the context is set to null.
- The base form of the nearest noun type token (e.g.: nouns, gerunds and acronyms).
- The nearest preposition.

In the case of the last two contexts, if there are no prepositions or nouns between the term and a verb, the context is set to the null context.

For each type of context pattern we count the number of context sets for each term. We use the Jaccard coefficient counted for the number of occurrences (CS) and a slightly modified version of this measure used in (Nenadic et al., 2006), CS$_2$. The context similarity between two terms t$_1$ and t$_2$ for a type of context is defined as follows:

$$CS_A(t_1,t_2) = \frac{|C_1 \cap C_2|}{A*|C_1 \cup C_2| + |C_1 \setminus C_2| + |C_2 \setminus C_1|}$$

where C$_1$ and C$_2$ are the sets of contexts defined respectively for t$_1$ and t$_2$ according to the chosen type of context patterns, A is equal to 1 or 2.

4.2 Coordination

The next type of syntactical information we use is the co-occurrence of terms in coordinated sequences. We took into account sequences of terms connected by conjunctions or commas but not with other types of tokens. All terms should be in the same grammatical case, and no additional tokens except terms, conjunctions and commas are allowed to be a part of a coordinated sequence.

The longest sequences of terms consisted of 8 medication names and chemical substances (components of medications). In our data we detect 1629 coordinated sequences of terms, that join 757 pairs of terms. The vast majority of them—564–occurred only once, but there are 4 pairs of terms, that occurred in coordinated phrases more than 100 times, e.g.: pancreas and spleen—335 times. As the differences in frequencies are extremely high, we decided to use a simple method of assigning the similarity of terms occurring in coordinated pairs based on the number of occurrences, see Tab. 1.

Table 1: Coordinated pairs similarity.

Frequency	Similarity	Numb. of pairs
1	0.25	564
2	0.30	94
3..9	0.50	73
10..49	0.75	15
50..99	0.90	7
> 100	1.00	4

4.3 Lexical Similarity

Terms that have the same head element usually describe related concepts, for example *lewa nerka prawidowa* ‘left kidney (is) normal’ and *nerka prawa* ‘kidney right’. If the head of two terms are the same then the head similarity for them is set to 1.

Terms that have common sets of modifiers are

Table 2: Different sets of coefficients.

1.	.05 .15 .15 .05 .05 .025 .025 .001 .001
	.01 .05 .05 .15 .02 .02
	.05 .05 .02

2.	all coeff. equal to 1/18
3.	.2 .1 .2 .1. 0 0 0 0 0 0 0 0 .1 0 0 0 0 0

X	.155 .155 .05 .05 .05 .025 .02 .001 .001
	.05 .05 .1 .02 .05 .01
	.143 .035 .035

also more related than those without any common modifier. As modifiers we understand all words included in the term except the head. For example the adjective *rwnomierny* ‘proportional’ indicates that the following terms are to a certain degree similar: *rwnomierne rozmieszczenie* ‘proportional distribution’ *rwnomierne gromadzenie* ‘proportional accumulation’. The lexical similarity between two terms is equal to the number of common modifiers divided by the number of modifiers of the longer term.

4.4 Overall Similarity

For each pair of terms, a final similarity was counted as a weighted sum of the 18 coefficients:

- neighboring left/right bases (2),
- neighboring left form (1),
- left/right POS contexts of length 2/3/4 (6),
- first left/right verb, noun, preposition (6),
- coordination coefficient (1),
- common head coeff. (1),
- common modifiers coeff. (1).

The initial intuition was that the most important features are direct left and right neighbors and verbal left neighbor, but several different coefficient combinations were tested to determine which features have the most positive influence on the results. Some of the tested schema are listed in Tab. 2.

At the beginning we chose two baselines: in the first one all coefficients are equally important (set number 2), in the second one a subset of only 5 features was chosen (3). Other sets were manually chosen according to our intuition about the role of the particular features (one example given as 1). The last set named ‘X’ was the result of adjusting coefficient values when comparing the results with manually prepared data (see sec. 5).

5 CLUSTERING

Automatic clustering was done using MultiDendrograms (Fernndez and Gmez, 2008) which implements several grouping strategies (single/complete linkage, unweighted/weighted average, unweighted/weighted centroid). Rough evaluation of the results showed the evident predominance of the variants in which similarity was counted on the basis of the unweighted average of coefficient values. The equilibrium between group size and their internal integrity was observed for results containing about 100 groups (a lot of groups were then singletons, but lowering the number of groups resulted in obtaining groups which contained non similar terms). In further experiments different weighting schemata of the coefficients used in the similarity measure were tested using the selected clustering strategy for the result set of about 100 groups.

The results of clustering were compared using the B-cubed measure (Bagga and Baldwin, 1998) positively evaluated in different experiments, e.g. (Amigó et al., 2009). This measure counts precision for every group element so it is sensitive both to presence and absence of the elements of groups. The results presented in Tab. 3 show some of the tested system configurations. Row and columns labels correspond to the combinations of weights given in Tab. 2 for both CS_1 and CS_2 definitions used for the first 15 features. Each cell of the table contains precision, recall and F-measure results when a model being a row label is taken as a reference one. The F-measure counted for all pairs of the models obtained was between 38% and 85% showing that the weights have real impact on the final grouping.

Table 3: Model comparisons.

a) CS_1 measure

	1	2	3
2	77.9/73.6/75.7	-	-
3	67.2/62/64.5	66.6/62/64.2	-
X1	77/75.4/76.2	71.5/73.4/72.5	68.7/71.7/70

b) CS_2 measure

	1	2	3
2	73.7/72.8/73.2	-	-
3	60.5/57.6/59	62.2/58.2/60.1	-
X2	70.5/73.2/71.8	65/65.2/65.1	66/68.2/67

To approximate the difficulty of the task, and to obtain evaluation data for the automatic clustering experiments, a manually prepared version of the partition was created. The 300 terms were grouped independently by two annotators. The only instruction given was to form large groups of terms which are of the same type (can represent concepts which are close in the *is-a* hierarchy). Two groupings contained 60 and 52 classes respectively. The distribution of the number of groups and their cardinality is given in Tab. 4.

Human annotators showed a tendency to create a few big groups. The first annotator tried to find as many relations as possible and created only 10 one elements groups, while the second annotator created several big groups but left the remaining terms unclustered—31 elements were left as one elements groups. The analysis of these two clusterings shows that the annotators grouped the terms according to different principles. The clustering of the first annotator more reflects the specific domain, while the second annotator tried to create more general groups according to more universal rules. For example, the first annotator created a group of urine features so terms denoting colour *barwa ta* 'yellow colour' and acidity *odczyn kwany* 'acidity' are in one group, while the second annotator left the term denoting colour in a separate group. The granularity of clusters is also different, e.g.: the first annotator created 2 groups of body parts and distinguished a group for parts of the urinary system, while the second created only one group of all body parts.

Table 4: Size and number of groups.

Gr	An. 1		An. 2		Join		Auto	
size	gr	el	gr	el	gr	el	gr	el
10..	7	153	8	184	6	93	3	50
5..9	10	70	7	47	10	65	16	101
2..4	24	67	14	38	41	110	42	110
1	13	10	31	31	32	32	39	39
total	54		60		89		100	
Max	53		63		24		24	

The resulting F-measure value of 66,5% showed that the task is really hard and the differences in grouping strategies can be quite high.

Next, a unified version of the classification was negotiated (column 'Join' in Tab. 4) and was used as a reference model for further experiments. One important rule in the process of creation joined classification was that if one annotator distinguished a subgroup of terms then annotators tried to divide big groups of terms into smaller ones. They also decided that the most important issue is adjusting a resource to the particular domain. The next decision was to take into account the same granularity of the related term groups. One of consequences of these decisions was distinguishing 13 groups of body parts. The final manual clustering contains more smaller groups of terms, and is a bit more similar to the clustering proposed by the first annotator (F-measure 70.8) than to the second one (F-measure 69.9%).

Table 5: Comparison of manual results.

	a_2	S
a_1	61.9/71.7/**66.5**	87.4/59.6/**70.8**
a_2	-	92.9/55.3/**69.4**

By adjusting the weights used in the definition of the similarity measure we obtained two models X1, X2 (for two similarity measures used respectively) which are most closely related to the manually obtained partition. The results shown in Tab. 6 are influenced by the difference in group numbers (89 groups vs 100).

The distribution of groups obtained by automatic clustering is given in Tab. 4 in column 'Auto'. A comparison with 'Join' clustering shows that there are very similar numbers of small groups (less than 5 elements), and the cardinality of the biggest group is the same. Manual inspection of the automatic grouping shows that many small groups are reasonably created even though some of them differ from the manual grouping. For example annotators created a group of blood types, in the 300 selected terms only type 'A' and 'B' is represented. In the automatic approach the term *krew* 'blood' was added to this group. Unfortunately, quite often, a completely unrelated term is added to a group of related terms. For example in the unified manual clustering a 4 element group containing terms denoting joints is distinguished. The same group is present in the automatic clustering, but additionally an unrelated term used in the data in the context of a bladder is added. i.e.: *nierwne ciany* 'uneven wall'. The biggest (24 elements) automatically created group contains 10 related terms from one 14 element manually created group. These terms describe diagnoses of the urinary system, the next 4 terms are to a certain degree related, but the remaining 10 elements are in that group by chance. The longest group fully consistent with manual clustering contains 6 elements. The manually created group contains an additional 3 elements.

Table 6: Best automatic models evaluation.

	X1	X2
a_1	63.8/33.9/44.2	64.9/33.1/43.9
a_2	66.4/34.4/45.3	68.1/33.9/45.3
S	**62/53.4/57.4**	**61.9/52.2/56.6**

6 CONCLUSIONS

Our results of automatic clustering show that in the case of a specific multi-topic domain text data for which no terminology or ontology resources are available, automatic clustering can be used to do preliminary term grouping. Even when only morpho-syntactic features are available, the results are of a quality good enough to be used as a starting point for further processing which may involve manual correction. The achieved F-measure of about 57% is not high, but in case of this task, which also proved difficult for well trained annotators, can be seen as good enough to be utilized in further domain ontology development. However, it turned out that morpho-syntactic information is not sufficient to build reliable clusters—additional sources of information will be searched for to improve quality of the results e.g. Wordnet which can be used to define semantic similarity between head elements of the different terms.

REFERENCES

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(5):613.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *LREC Workshop on Linguistics Coreference*, pages 563–566.

Baneyx, A., Charlet, J., and Jaulent, M.-C. (2006). Methodology to build medical ontology from textual resources. *AMIA Annual Symposium proceedings*, 2006:21–25.

Cimiano, P. (2006) Ontology Lerning and Population from Text. pages 85–184. Springer.

Fernndez, A. and Gmez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, 25:43–65.

Frantzi, K., Ananiadou, S., and Mima,2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *Int. Journal on Digital Libraries*, 3:115–130.

Ittoo, A. and Maruster, L. (2009). Ensemble similarity measures for clustering terms. In *Congres on Computer Science and Information Engineering*, volume 4, pages 315–319.

Le Moigno, S., Charlet, J., Bourigault, D., Degoulet, P., and Jaulent, M.-C. (2002). Terminology extraction from text to build an ontology in surgical intensive care. In *Proceedings of the Workshop Machine Learning and Natural Language Processing for Ontology Engineering*.

Lin, D. and Pantel, P. (2001). Induction of semantic classes from natural language text. In *KDD'01*, pages 317–322.

Navigli, R., Velardi, P., and Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, 18(1):22 – 31.

Nenadić, G., Spasić, I., and Ananiadou, S. (2004). Automatic discovery of term similarities using pattern mining. *Int. Journal of Terminology*, 10(1):55–80.

Nenadic, G., Spasic, I., and Ananiadou, S. (2006). Term clustering using a corpus-based similarity measure. In Sojka, P., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 2448 of *LNCS*, pages 89–109. Budapest, Hungary.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Ushioda, A. (1996). Hierarchical clustering of words. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 1159–1162, Stroudsburg, PA, USA. ACL.

Woliski, M. (2006). Morfeusz—a Practical Tool for the Morphological Analysis of Polish. In Kopotek, M., Wierzcho, S., and Trojanowski, K., eds, *Intelligent Information Processing and Web Mining, IIS:IIPWM'06*, pages 503–512. Springer.