Combining Wordnet and Morphosyntactic Information in Terminology Clustering

Agnieszka Mykowiecka   Małgorzata Marciniak
Institute of Computer Science, PAS, Jana Kazimierza 5, 01-248 Warsaw, Poland
agn@ipipan.waw.pl, mm@ipipan.waw.pl

ABSTRACT
The paper presents results of clustering terms extracted from economic articles in Polish Wikipedia. First, we describe the method of automatic term extraction supported by linguistic knowledge. Then, we define different types of term similarities used in the clustering experiment. Term similarities are based on Polish Wordnet and morphosyntactic analysis of data. The latter takes into account: term contexts, coordinated sequences of terms, syntactic patterns in which terms appear and words that are parts of terms (such as their heads and modifiers). Then we performed several experiments with hierarchical clustering of the 400 most frequent terms. We present the results of clustering when different groups of similarity coefficients are applied. Finally, we present an evaluation that compares the results with manually obtained groups. Our results prove that morphosyntactic information can help or even serve themselves for initial clustering of terms in semantically coherent groups.

KEYWORDS: terminology extraction, terminology clustering, Polish.

Proceedings of COLING 2012: Technical Papers, pages 1951–1962,
COLING 2012, Mumbai, December 2012.

1 Introduction

Many NLP applications, like text indexing, information extraction or question answering, rely on sets of predefined concepts which can be identified within texts, on a given subject. Such concepts form either hierarchical ontologies or flat terminology lists. Although there are already very many works aimed at making these kinds of resources available, still, the existing data is very limited with regard to the chosen application/knowledge domain as well as natural language addressed. An overview of existing approaches to automatic terminology extraction is included in (Pazienza et al., 2005) while problems with ontology creation are described, among others, in (Cimiano, 2006).

For some NLP purposes flat terminology lists are sufficient, but for others, like IE, ontologies representing relations between particular terms are more adequate. Unfortunately, domain ontologies' availability is very limited and what is even more crucial – they are rarely adequate for the purpose at hand. They are either too specific or too general, or do not cover the appropriate domain, or they are outdated. Ontology reuse and projection is still a very difficult and unresolved issue. As a result, when general ontologies like SUMO (Pease and Niles, 2001) are insufficient, or are not available in a particular language, a new dedicated ontology has to be created. In such a case, a list of concepts to be included in the ontology can be prepared manually by a domain expert, or can be (at least initially) extracted from texts similar to those which are to be processed.

In this paper we address the problem of finding interesting concepts in Polish economic texts and organizing them in coherent groups which can be further analyzed more easily than a long unstructured term list. The identified sets of terms will be used for developing a domain model which will constitute the base of an information extraction system. The basic idea is to facilitate the construction of IE systems by automation of the initial stage of domain model creation, i.e. defining concepts which can be addressed in the selected type of texts, and relate them to particular language expressions. The same method may be used later on to gather new concepts which appear in time and relate them to the already existing ones on the basis of the contexts of their occurrences.

As a test domain we have chosen economy, in particular, economic articles of Polish Wikipedia. In the paper we present the process of selecting term candidates, their ordering according to the defined importance measure and clustering. To make the approach usable for many domains we assume that only general language resources like a morphological tagger and Wordnet are used.

2 Data

The experiment was conducted on the economic articles taken from the Polish Wikipedia. Only textual content of these articles was taken into account. The data was collected in 2011 and contains 1219 articles that have economics related headings and articles linked to them. The data contains about 456,000 tokens. An initial linguistic analysis of the plain texts was performed. It consisted of the following steps:

- Segmentation into tokens. We distinguish words (365,042), numbers (14,906) and punctuation marks (76,239).
- 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 publicly available Polish POS tagger Pantera (Acedański, 2010) that cooperates with a general-purpose morphological analyzer of Polish Morfeusz SGJP (Woliński, 2006). The Pantera tagger allows us to define a separate dictionary in which we can describe unknown tokens. So we defined an additional domain dictionary containing 741 entries of word-forms (not recognized by the Morfeusz analyzer). Many of them have more than one morphological characterization, e.g. *podsektor* 'subsector' that represents a noun in the nominative or accusative case. In case of adjectives there are usually more possible interpretations, for example the word-form *proekologiczne* 'proecological' has 7 different characterizations in the additional dictionary. Our dictionary does not describe all unknown tokens as we decided not to analyze foreign words and proper names that are not present in Morfeusz. In the data, many notions are translated into foreign languages, e.g. *Spółdzielnia europejska* 'European Cooperative Society' has in the data its Latin equivalent 'Societas Cooperativa Europaea', and all three tokens are annotated with *ign* tag that indicates an unknown word. So still 10,796 tokens have no morphological characterization.

- Improving tagger results. We defined 84 rules in Spejd (Przepiórkowski, 2008) (a cascade of regular grammars) in order to correct Pantera decisions, and to extend some descriptions. The advantage of using this method is the possibility of taking contexts into account. Spejd rules are particularly helpful in correcting some regular tagging errors in frequently occurring phrases. They corrected or extended over 4,000 token descriptions. The changes in the tagset consisted in the introduction of the *number* tag assigned to Arabic as well as Roman numerals, and extending descriptions of abbreviations (*brev* POS). In Morfeusz, an abbreviation is characterised only by a specification whether it has to be followed by a full stop. We extended its description with information about the type of word or phrase it abbreviates to allow for constructing correct grammatical phrases containing the abbreviation.
- Removing improperly recognized sentence endings after abbreviations.

3 Terms identification

The very first problem while doing terminology extraction is to define what a domain term really is. Unfortunately, there exists no strict definition of this concept and usually only pragmatical approaches are taken. For the purpose of this work we defined a term as a noun phrase which occurs more frequently in domain specific texts than in the general language. Thus, we decided not to follow the approach in which linguistic information is neglected (like (Wermter and Hahn, 2005)), but to use morphological information at the stage of candidate selection. We also use this information while defining similarity coefficients.

For term candidates we choose noun phrases of a limited internal complexity, i.e. we assume that they are built according to one of the following syntactic schemata of which the first four are very frequent and the last one is much less common. In particular, we do not allow for prepositional phrases to occur within the terms (compare the results of the terminology extraction task described in (Marciniak and Mykowiecka, 2012)).

- a single noun or an abbreviation, e.g. *bank* 'bank', *EC* 'European Commission';
- a noun followed (or, more rarely, preceded or surrounded) by an adjective, e.g. $administracja_n$ $publiczna_{adj}$ 'public administration', $wysoki_{adj}$ $dochód_n$ 'high income', $ogólna_{adj}$ $sytuacja_n$ $gospodarcza_{adj}$ 'general economic situation';

- a noun followed by another noun in genitive, e.g. $kurs_{n,nom}$ $walut_{n,gen}$ ‘exchange rate’;
- a combination of the last two structures, e.g. $walne_{adj,nom}$ $zgromadzenie_{n,nom}$ $akcjonariuszy_{n,gen}$ ‘general meeting of shareholders’,
- a noun preceded by an adjectival phrase, e.g. *wschodnia i południowa Afryka* ‘East and South Africa’.

For recognizing the selected types of nominal phrases, a cascade of six simple shallow grammars was created. Its rules operate on the results of morphological analysis described in section 2. The gradual phrase creation starts with adjective modifiers and then genitive modifiers are added.

To make our data a little more coherent (domain related) we eliminated time related expressions (names of months, nouns like ‘hour’, ‘minute’, adjectives like ‘late’) which are studied separately. We also excluded selected sets of nouns and adjectives which can be thought of as a kind of ‘stop words’ for the terminology extraction task, that is words which can be used in very many contexts and which themselves do not constitute terms elements. These are adjectives like *dany* ‘given’ or pronouns. The list was built up in our previous terminology extraction experiment (Mykowiecka and Marciniak, 2012) and supplemented with new elements in the current one. The list which contains 65 words is used only additionally, many such phrases can be eliminated at the later stage of ordering term candidates.

Applying the adopted set of rules to the data resulted in obtaining 80,212 types of phrases in which there are 45,144 top level types occurring 104,576 times. The longest (non overlapping) phrases which can be built starting from subsequent text positions were extracted. Their internal structure was annotated by markers showing subphrase boundaries. For the resulting set of phrases, we performed an analysis similar to that proposed in (Frantzi et al., 2000). In this approach both the entire high-level phrases and the internal nominal subphrases are taken into account, e.g. in the phrase *rzecznik dyscypliny finansów publicznych* ‘advocate for public finance discipline’ we also encounter the subphrases *dyscyplina finansów publicznych* ‘public finance discipline’ and *finanse publiczne* ‘public finance’. Taking subphrases into account is important, as for example, the following phrases: *kapitał obrotowy* ‘working capital’ *system emerytalny* ‘pension system’ and *akt notarialny* ‘notarial deed’ did not occur in isolation in the data.

As Polish is an inflectional language, phrases which are identified within the text are of different forms (e.g. $\text{kurs}_{n,acc}$ $\text{walut}_{n,gen}$, $\text{kursie}_{n,loc}$ $\text{walut}_{n,gen}$ ‘exchange rate’) so the usual processing stages like counting phrase frequencies and preparing a list of phrase types became difficult. To overcome this problem we produce an artificial base form of every identified phrase occurrence, by taking base forms assigned by the tagger to its elements, i.e. $\text{kurs}_{n,nom}$ $\text{waluta}_{n,nom}$.

All extracted phrases are ranked according to the value of a specially defined coefficient (C-value) which is calculated basing on the occurrences of the phrase in the text as a stand alone phrase and its occurrences within other phrases from the list. This allows us to identify terms which never (or very rarely) occur in isolation, and to some extent, to filter out erroneous phrases (those which are recognized by a shallow grammar as one phrase but in fact are incomplete although grammatically sound) or are built up of more than one phrase (like *zamieszkania właściwość różnych organów* ‘living jurisdiction of different authorities’ which resulted from *podlegają z uwagi na swe miejsce zamieszkania właściwości różnych organów* ‘fall under jurisdiction of different authorities depending on their place of living’).

We used a slightly modified definition of C-value which is given below. p – is a phrase under consideration, LP – is a set of phrases containing p, and P(LP) – the number of types of phrases differing in elements which are adjacent to p, that is the sum of different direct left and right one-word contexts counted separately (e.g. if the phrase *angielski bank* ‘English bank’ occurs in three types of longer phrases: *angielski bank inwestycyjny* ‘English investment bank’, *najstarszy angielski bank inwestycyjny* ‘the oldest English investment bank’ and *bankructwo najstarszego angielskiego banku inwestycyjnego* ‘bankruptcy of the oldest English investment bank’ , P(LP) is set to 2).

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

where $lc(p) = log_2(length(p))$ if $length(p) > 1$ and 0.1 otherwise;

To eliminate phrases which are not from the economy domain, but occur in all types of texts similarly often, we compared the list of phrases obtained for Wikipedia economic texts with phrases obtained form the balanced one million word subcorpus of NKJP (the corpus of general Polish (Przepiórkowski et al., 2012)) using the same processing schema. Table 1 shows how many terms are recognized in both corpora and how many of them have a grater C-value in each data set. Less than 10% of terms recognized in economic texts are also recognized in NKJP data — the longest common phrases have 5 words.

Table 1: Comparison with general corpus

Terms	common	C-value greater in econom.	C-value greater in NKJP
1-word	4089	767	3322
2-words	2558	1133	1425
3-5-words	201	98	103
Total	7848	1998	4850

There are a number of phrases with a greater C-value for NKJP subcorpus and relevant to the economic domain, e.g. *skarb państwa* ‘state treasury’, *urząd skarbowy* ‘treasury office’, *ustawa budżetowa* ‘budget act’. So we decided that phrases which have a greater C-value counted in the context of general texts than that counted for economic data, should be manually inspected.

For the clustering experiment, the first 400 terms from the list, ranked according to the C-value coefficient, were chosen. On the bases of the comparison with NKJP terms, from the original list we removed one-word terms like: *grupa* ‘group’, *przyklad* ‘example’, *funkcja* ‘function’; and a few multi-word terms, e.g: *wszcząć postępowanie* ‘initiation of proceedings’, *różny rodzaj* ‘different types’. Removed terms were substituted with the subsequent terms from the list. Choosing a relatively small number of terms was motivated by the need for manual checking of the results.

4 Defining similarity features

At the next stage of domain model creation, a list of terms is organized into clusters which group elements addressing similar concepts. This is most frequently done manually by domain experts, but manual processing of a long list of names is time consuming and prone to errors. To perform this task automatically it is necessary to decide how to represent term similarity. In our approach we decided to use morphosyntactic information (in this case we follow the ideas presented in (Nenadić et al., 2004)), as well as information included in Polish Wordnet.

4.1 Contextual similarity

Contextual similarity is based on the contexts in which terms appear. We consider left and right contexts of terms separately. Contexts are not allowed to cross sentence or paragraph boundaries. We decided to consider the following types of context patterns:

- 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 the null context.
- The base form of the nearest noun type token (e.g. nouns, gerunds).
- 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.

4.2 Coordination

Co-occurrence of terms in coordinated sequences is the next type of information we take into account when finding similar terms. We find sequences of terms connected by conjunctions or commas. All terms should be in the same grammatical case, and can be preceded by a preposition. The following example of a coordinated sequence: *<akcje>, <obligacje> i <instrumenty pochodne>* '<shares>, <bonds> and <derivatives>' joins terms denoting various financial instruments. We also consider coordination of prepositional phrases that consist of a preposition and a term, where terms are in the same grammatical case. See an example of such a phrase: *dla <osoby prywatnej> i dla <jednostki organizacyjnej>* 'for <a private person> or for an <organization unit>'. We do not check the wordforms of prepositions in coordinated sequences, but the grammatical case of terms has to be the same. This is a rough method of coordinated terms recognition and needs further refinements. For example, it will not recognize the following coordination: *eksportowane do Niemiec*$_{gen}$ *i na Litwę*$_{acc}$ 'exported to Germany and to Lithuania'. However, it excludes the majority of cases where two terms preceded by prepositions are separated by a comma and they belong to two different parts of a sentence, e.g: *W <Polsce>*$_{loc}$*, mimo <wpisania pojęcia konsumenta>*$_{gen}$ *do konstytucji ...* 'In [Poland], despite of [entering the notion of a consumer] into the constitution ...'.

In our data we detected 5,885 coordinated sequences of terms, which join 9,807 different pairs of terms. The vast majority of them occurred only a few times, only 9 pairs of terms occurred in coordinated sequences more than 10 times. The most frequent pair of terms *<towar>* 'product' and *<usługa>* 'service' occurred 74 times. For the selected 400 terms, 157 coordination pairs were found within the texts.

4.3 Syntactic patterns

Besides the coordination sequences we recognize several syntactic patterns that indicate similarity between terms. These patterns contain the following Polish phrases/words: *taki jak* 'such as', *czyli* 'or, that is', *na przykład* 'for example', *to jest* 'that is' and *zarówno...jak i* 'both...and also' and their equivalents. The first four patterns have the following construction:

<term1> [key phrase] <list of terms>

while the last one has slightly different form:

[key phrase 1] <term1> [key phrase 2] <list of terms>.

In the above patterns <list of terms> is the coordination of terms with limitations and internal similarity measures described in 4.2. These constructions recognize similarity between pairs built up from <term1> and all terms in the <list of terms>. Let us consider the following example:

wiele cech <oferty rynkowej> takich jak <cena>, <jakość> i <forma płatności>

'many features of <market offer> such as <price>, <quality> and <form of payment>'

The above phrase indicates that following 3 pairs of terms are similar:

- *<oferty rynkowej>* 'market offer' and *<cena>* 'price'
- *<oferty rynkowej>* 'market offer' and *<jakość>* 'quality'
- *<oferty rynkowej>* 'market offer' and *<forma płatności>* 'form of payment'

In the data we detected 545 pairs of similar terms recognized by the above lexical patterns from which 85 are used in the clustering experiment of 400 terms.

4.4 Lexical Similarity

Terms that have the same head element usually describe related concepts, for example *kurs obcej waluty* 'foreign currency exchange rate' and *kurs dolara* 'dollar exchange rate' have the same head element *kurs* 'exchange rate', and describe similar notions. In the task we promote terms with the same head. If the heads of two terms are the same, then the head similarity coefficient for these phrases is set to 1. We do not consider different meanings of heads so the following phrases: *klasa robotnicza* 'working class' and *klasa szkolna* 'classroom' are set to 1.

Terms that have common words are also more related than those without any common words. Counting them we exclude common heads. For example the common adjective *budżetowy* 'budgetary' indicates that the following terms are to a certain degree similar: *dotacja budżetowa* 'budget subsidy' *wydatek budżetowy* 'budget expenditure' and *założenia budżetowe* 'budget assumption' To establish these types of similarities for all term pairs we counted how many common words they have (except common head elements). The similarity between two terms is equal to the number of common modifiers divided by the number of modifiers of the longer term.

4.5 Wordnet similarity

Polish Wordnet (PlWordNet, (Piasecki et al., 2009)) is one of the biggest resources of the type introduced by Princeton Wordnet (Miller, 1995), but it mainly describes general language and it contains mostly one word items. Domain terminology usually contains a prevalence of multiword expressions. On our list, among 400 terms, 130 are one word expressions and 270 are longer. All one word terms have at least one sense defined in PlWordNet. For multiword expressions the situation is drastically different: 52 phrases are defined in PlWordNet while 218 phrases are not. For 3 of them their head elements are also not described: *Brytania* 'Britain, *środki* 'resources', *Adam* 'Adam' (two of them are proper names, for the word 'resources' only the singular form is defined which has different meanings).

The above statistics show that in order to utilize information given in PlWordnet, operating only on information in the phrases which appear within the data, is insufficient. Thus, we decided to calculate the similarity between terms on the basis of information given both on the terms themselves and on their head elements. The schema of calculating similarity between terms A and B was defined in two steps. In the first one an initial similarity value is set to:

- if both terms appear in PlWordNet and share at least one synset — 1/minimum of synsets defined for A and B;
- otherwise, if A appears in PlwordNet and belongs to the same synset as one of the B hiper- or hiponims – 0.5/number of synsets of A;
- otherwise, if B appears in PlwordNet and belongs to the same synset as one of the B hiper- or hiponims – 0.5/number of synsets of B;

In the second step, when at least one of the terms is longer than one word, the similarity value is assigned to a minimum from the number resulting from the following additions and 1:

- if A is a multiword term:
  - if A's head belongs to at least one synset to which B also belongs: +0.25;
  - otherwise, if A's head belongs to at least one synset to which the head of a multiword A also belongs: +0.1;
  - if B belongs to the same synset as a hiper- or hiponim of A's head: +0.15
- if B is a multiword term
  - if B's head belongs to at least one synset to which any hiper- or hiponim of B also belongs: +0.05
  - if (oneword) A belongs to the same synset as the head of B: +0.2
  - if (oneword) A belongs to the same synset as a hiper- or hiponim of B: +0.1

This process resulted in 298 nonzero coefficients. 19 pairs were judged to be equivalent (similarity 1), e.g. *<dochód>-<zysk>* 'income-gain', *<prawo>-<zasada>* 'law-rule. One example was incorrect: *<model>-<klient>* 'model-customer'.

4.6 Overall Similarity

All similarity tables were rescaled in such a way that the highest coefficient for each measure is equal 1. In all experiments described below, an overall similarity of a pair of terms was calculated as a weighted sum of up to 19 coefficients:

- neighboring left/right form (lf, rf),
- left/right POS contexts of length 2/3/4 (c2l, c3l, c4l, c2r,c3r,c4r),
- first left/right verb, noun, preposition (l_v, l_n, l_p, r_v, r_v, r_p),
- coordination coefficient (crd),
- syntactic patterns coefficient (syn)
- common head coeff. (head),
- common modifiers coeff. (mod),
- wordnet similarity (wdnet).

5 Clustering

Automatic clustering was done using MultiDendrograms (Fernández and Gómez, 2008) performing hierarchical clustering. From several options, the unweighted average of similarity coefficient values was selected on the basis of the results of the preliminary tests.

As no resource which can be used as a reference set exists, to enable the evaluation of the results, a manually prepared version of the partition of 400 terms was created. The only instruction given to a person doing this task was to group similar elements even if they cannot be treated as subtypes of one concept. The result, which was verified by the second annotator, comprises 127 group, of which 30 contain only one element. The maximal group size is 14.

In the experiments, different weighting schemata of the coefficients used to calculate the overall similarity measure were tested. The exact values of the weights assigned for some selected models are given in Table 2. To check the impact of morphosyntactic features on the result obtained while using only PlWordNet data, automatic clustering was done for the models belonging to the three groups described below.

- only Wordnet similarity (as defined above) has nonzero (i.e. 1) weight – model W1,
- all weights are non zero — models W4, W5 and W6,
- Wordnet similarity is assigned 0, its weight is distributed among other weights which are initially set as in W4 – W2.

Table 2: The selected models characterization

	fl	fr	syn	c2l	c2r	c3l	c3r	c4l	c4r	crd	mod	head	r_v	r_n	r_p	l_v	l_p	l_n	wdnet
w2	.11	.11	.058	.058	.058	.033	.033	.013	.013	.058	.058	.108	.058	.058	.018	.058	.043	.053	.00
w4	.10	.10	.050	.050	.050	.025	.025	.005	.005	.050	.050	.100	.050	.050	.010	.050	.035	.045	.15
w5	.13	.13	.100	.015	.020	.010	.010	.003	.002	.050	.050	.080	.040	.050	.050	.100	.030	.030	.10
w6	.10	.10	.050	.050	.050	.025	.020	.001	.001	.050	.050	.100	.020	.050	.010	.123	.040	.040	.13

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 to both – presence and absence of the elements of groups. The results presented in Tab. 3 show some of the tested system configurations. Rows correspond to the combinations of weights given in Tab. 2. Each cell of the table contains precision, recall and F-measure results obtained when comparing the models to the manual clustering (rescaled into the 0-100 range).

Table 3: Model comparisons with manual grouping.

	127 groups	best F-value	nb of groups
W1	64.2/74.1/68,8	84.9/71/77.4	175
W2	62.4/60.9/61.7	82,5/56,9/68,2	205
W4	74.3/73.9/74.1	94.9/69.6/80.3	190
W5	76.8/73.6/75.1	90.6/69.6/78,7	175
W6	76.4/73.0/73.8	94.4/68.8/79.6	191

By adjusting the weights used in the definition of the similarity measure, we obtained an enhancement of the clusters matching which did not vary much (for reasonable weight distribution). For all these models the results were about 5% better than those obtained using only Wordnet data. Using morphosyntactic information alone also gave usable results at the level of about 62%.

Conclusions

In the paper we presented the results of the process of detecting coherent groups within terminological phrases extracted from real texts from the economy domian. The obtained results show that in the case where semantic information is lacking, morphosyntactic description of the contexts of term occurrences can help in a terminology clustering task. Even when only morphosyntactic features are available, the results achieved can make further manual clustering much easier. However, adding other sources of information, like Wordnet relation for phrase head elements, improves the results.

The F-measure of about 75% achieved when comparing the automatically obtained clusters to manually obtained groups is not high, but in the 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. The presented method can be used for texts in any domain or language but the quality of the results highly depends on the quality of lexical tools used for preprocessing. Our results of the lexical preprocessing stages showed that the quality and coverage of Polish taggers are not very good when dealing with more specific texts. In such cases even a small additional dictionary might be necessary to obtain good results. On the other hand, a big common part which economic texts have with general language used in newspapers make the terminology selection stage less precise.

References

Acedański, S. (2010). A morphosyntactic Brill tagger for inflectional languages. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S., editors, *Advances in Natural Language Processing*, volume 6233, pages 3–14. Springer Berlin / Heidelberg.

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 *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Cimiano, P. (2006). *Ontology learning and population from text. Algorithms, evaluation and applications*. Springer.

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

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

Marciniak, M. and Mykowiecka, A. (2012). Terminology extraction from domain texts in Polish. In Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M., editors, Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions, Springer (in press)

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mykowiecka, A. and Marciniak, M. (2012). Clustering of medical terms based on morphosyntactic features. In Proc. of International Conference on Knowledge Engineering and Ontology Development KEOD 2012, Barcelona. SciTePress Digital Library..

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

Pazienza, M., Pennacchiotti, M., and Zanzotto, F. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In Sirmakessis, S., editor, *Knowledge Mining Series: Studies in Fuzziness and Soft Computing*. Springer Verlag.

Pease, A. and Niles, I. (2001). Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Wermter, J. and Hahn, U. (2005). Massive biomedical term discovery. In *Discovery Science, LNCS 3735*, pages 281–293.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Kłopotek, M., Wierzchoń, S., and Trojanowski, K., editors, *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, pages 503–512. Springer.