Collaborative Approach to WordNet and Wikipedia Integration

Julian Szymański, Rafał Korytkowski, Henryk Krawczyk
Faculty of Electronics, Telecommunications and Informatics
Department of Computer Architecture
Gdańsk University of Technology, Poland

Email: julian.szymanski@eti.pg.gda.pl, rafal.korytkowski@eti.pg.gda.pl, henryk.krawczyk@eti.pg.gda.pl

***Abstract*—In this article, we present a collaborative approach to creating mappings between WordNet and Wikipedia. Wikipedia articles have been first matched with WordNet synsets in an automatic way. Then, such associations have been evaluated and complemented in a collaborative way using a web application. We describe algorithms used for creating automatic mappings as well as a system for their collaborative development. The outcome enables further integration of WordNet and Wikipedia, which can be used in Natural Language Processing algorithms.**

***Index Terms*—WordNet Wikipedia integration, ontology matching, information retrieval, text representation, natural language processing**

I. Introduction

In today's world, text is the main medium for presenting and exchanging information. According to Royal Pingdom [1] a company that monitors the Web, in 2010 people sent 107 trillion e-mails, 25 billion tweets (Short messages shared via http://twitter.com/), existed 255 million websites and 152 million blogs. Most of such resources are unstructured, thus they are very difficult to process by the computers. At the same time more and more effort is put into developing technologies, which may help processing and extracting knowledge from that overwhelming amount of information automatically.

The Semantic Web [2] is an idea aiming at extending the Web with meta data to support the automatic processing of its content. Typically semantic is introduced by annotating words, pages or other Web resources with references to ontologies [3]. For that to be possible, ontologies need to contain tremendous amount of structuralized information and instantly evolve with the culture and language. It can be only achieved with at least partial automation of their construction. It is an interdisciplinary endeavor engaging such fields as data mining, natural language processing or artificial intelligence and cognitive sciences [4].

The goal of this paper is to present a way to integrate existing linguistic databases to satisfy the need for a robust ontology. In particular, a mapping between WordNet dictionary and Wikipedia will be created. The databases were chosen due to their extensive usage in Natural Language Processing tools [5], however, presented approach is equally applicable to other resources.

Since it is not possible to create accurate mappings entirely automatically, a collaborative approach for evaluation and improvement of automatic mappings has been used. It allows to engage many people in evaluation of automatically created mappings and manual construction of additional mappings.

The paper is constructed as follows: in sections II-IV we describe Wordnet and Wikipedia repositories and the work related to integration of their resources. The next section describes the way for pruning Wikipedia data. In section VI we describe our method that automatize process of creating mappings between Wordnet synsets and Wikipedia articles. In next section we provide results of collaborative evaluation. The conclusions and future work has been proposed in the last paragraph.

II. WordNet

WordNet is a lexical database of English language [6]. It was originally developed and is maintained at Princeton University.It is both a dictionary and thesaurus. It contains nouns, verbs, adjectives and adverbs that are arranged in sets of synonyms called *synsets*. Each synset represents a unique word meaning and has its own definition. For example word horse has five meanings:

- **horse, Equus caballus** *[solid-hoofed herbivorous quadruped domesticated since prehistoric times]*
- **horse, gymnastic horse** *[a padded gymnastic apparatus on legs]*
- **cavalry, horse cavalry** *[horse troops trained to fight on horseback]*
- **sawhorse, horse, sawbuck, buck** *[a framework for holding wood that is being sawed]*
- **knight, horse** *[a chessman shaped to resemble the head of a horse; can move two squares horizontally and one vertically (or vice versa)]*

The synsets are linked together forming a semantic network. Links between synsets are considered the most valuable asset of WordNet. They represent semantic and lexical relationships between different word meanings.

The database in its current 3.0 version contains 155,287 words arranged in 117,659 synsets and 206,941 pairs word-synset (senses). The number of links between synsets amounts to 243,229.

The most widely implemented relations between synsets are:

 ISBN: 978-1-61208-206-6

TABLE I
RELATIONS IN WORDNET

	Nouns	Verbs	Adjectives	Adverbs
Hyponymy/hypernymy	84,427	13,239	-	-
Meronymy/holonymy	22,187	-	-	-
Similarity	-	-	21,386	-
Antonymy	2,152	1,093	4,024	710
Other	86,777	50,575	41,486	3,334
Total	111,766	64,955	62,872	4,044

TABLE II
PHRASES PER SYNSETS

Phrases	Nouns	Verbs	Adjectives	Adverbs
1	42,054 (51%)	8,041 (58%)	11,353 (63%)	2,400 (66%)
2	25,780 (31%)	3,146 (23%)	4,217 (23%)	771 (21%)
3	8,674 (11%)	1,280 (9%)	1,435 (8%)	289 (8%)
4	3,359 (4%)	623 (5%)	595 (3%)	91 (3%)
>= 5	2,248 (3%)	677 (5%)	556 (3%)	70 (2%)

- Hyponyms and hypernyms. A hyponym shares a type-of relationship with its hypernym. For instance *cat* is a hyponym of *wildcat* or *wildcat* is a hypernym of *cat*. Hyponyms and hypernyms have a common root and are transitive. For example if *wildcat* is a hyponym of *tiger cat* then *cat* is a hyponym of *tiger cat* as well.
- Meronyms and holonyms. A meronym shares a part-of relationship with its holonym. For instance *roof* is a meronym of *building* or *building* is a holonym of *roof*. Such relationships are not always transitive and have been divided into six types: component - object (*branch* - *tree*), member - collection (*tree* - *forest*), stuff - object (*aluminium* - *airplane*), portion - mass (*slice* - *cake*), feature - activity (*paying* - *shopping*), place - area (*Princeton* - *New Jersey*) [7].
- Antonym is a relationship between two synsets having opposite meanings, which may be defined for nouns, verbs, adjectives and adverbs such as *work* - *idle*, *ugly* - *beautiful*, *cold* - *hot*.
- Troponym is a relationship between synsets of two verbs with a different intensity of a certain property such as *like* - *love* (by the intensity of emotions), *sip* - *drink* (by the speed of consumption).

Beside relations between synsets belonging to the same part of speech, there are morphosemantic relations, which combine words with the same root such as *assistant* (noun) - *assist* (verb) - *assistive* (adjective).

Another important factor that will be used later in this paper is a number of phrases per synset (Table II). It can be observed that over half of all synsets define only one phrase.

III. WIKIPEDIA

Wikipedia does not need much introduction. It is among ten of the most visited websites on the Internet according to [8]. The project started in 2001. Its aim was to create the biggest and open encyclopedia in the world. It has also revealed a phenomena of collaborative work. Over 10 years its users have created 20 million articles in 268 languages.

Wikipedia uses a concept of an article as the atom of knowledge. An article must conform to a few rules defined in the Wikipedia Manual of Style ,which are easy to present using an excerpt of an article, e.g.:

> **Horse** *[The horse (Equus ferus caballus) is one of two extant subspecies of* Equus ferus, *or the wild horse. It is a single-hooved (ungulate) mammal belonging to the taxonomic family Equidae. The horse has evolved over the past 45 to 55 million years from a small multi-toed creature into the large, single-toed animal of today.]*

Titles of articles must be unique, thus if a word has more meanings, a title is usually concatenated with an additional expression in parenthesis. For instance there are articles titled as *Horse (Equus ferus caballus)*, *Horse (gymnastics)*, *Horse (geology)*, etc.

Different meanings of a word can be found through disambiguation pages. They are special articles, which contain links to different meanings and can be easily recognized as they belong to a special category, use a certain template or have the *(disambiguation)* keyword in their titles. In general other meanings of a word can be also found at the top of an article and they are preceded with *For other uses, see....*

It is also important to note that *Horse (gymnastics)* is not an article, but a redirect to the *Vault (gymnastics)* article. Redirects can be synonyms, but also plural forms or misspellings. If we are redirected to a page, we will see *Redirected from...* at the top.

In addition, *Horse (Equus ferus caballus)* is assigned to 17 categories such as *Animal-powered transport*, *Domesticated animals*, *Equus*, or *Horses*. Categories form a hierarchical structure of Wikipedia. They are not articles, but special entities, which contain a short description and a link to a related article. For instance, *Equus* links to the *Equus (genus)* article. Categories are linked together and can be represented as a graph. Both articles and categories may belong to many other categories. In rare cases we may experience cycles while traversing the graph.

Hiperlinks may refer to different sections of an article, other articles or outer pages. There is a measure of the number of links pointing to a certain article, which stands for its popularity. Hiperlinks may also be used as a supplement for redirects to find different synonyms.

Links to other languages are a particularly interesting aspect of Wikipedia. It is a unique among other encyclopedia's property, which can be used to translate terms.

Wikipedia, in contrast to WordNet, covers much more knowledge. There are 3.8 million articles in English with 77.1 million internal links and 5.2 million redirects according to Wikimedia statistics [9]. On the other hand Wikipedia is less organized and more erroneous than WordNet.

IV. RELATED WORK

In order to integrate different linguistic databases common terms between them need to be found. Ruiz-Casado et al. in their work [10] tag Wikipedia articles with WordNet synsets.

 ISBN: 978-1-61208-206-6

They use Simple Wikipedia, which is a version designated for people learning English with less articles and using only basic vocabulary. In their approach, they apply a disambiguation algorithm based on the Vector Space Model to determine similarity between an article and a synset. They ran the algorithm against 1,841 articles, 33% of which were not matched with WordNet synsets, 34% were matched with exactly one synset and 33% required disambiguation. In case of articles, which did not require disambiguation the accuracy was 98% and 84% in the other case.

The reported results were satisfactory; however, we did not expect to come close to that level when applying the algorithm to the full Wikipedia, because of a significant difference in the number of articles and their complexity. Therefore, we decided to take a different path.

Another approach to the automatic integration of Wikipedia and WordNet has been based on the word co-occurrence analysis. The analysis is performed between a synset definition and a first paragraph of a Wikipedia article [11]. The obtained results (39.51% and 49.28% quality depending on the method) evaluated for 500 test mappings indicate the method can be useful, but the method requires contribution of humans.

The next approach called YAGO is an ontology constructed using Wikipedia and WordNet [12]. Text mining algorithms from those resources allow to extract over 2 millions of objects and 20 millions of related facts. The project managed to construct around 15,000 direct mappings between WordNet synsets and Wikipedia articles in an automatic way [13].

WordNet is developed as a research project in a closed academic environment. The first version of the dictionary appeared in 1993, and now a third version is available. The dictionary is publicly available, but its modification is restricted from internauts. Probably, the reason for that, is the fact that the lexicon is organized as a set of text files in a specific format, which makes it hard to apply cooperative approach for WordNet development. The lack of cooperative editing functionality is the biggest barrier to scale-up a semantic database.

In our research, we develop the WordVenture portal [14], which provides mechanisms for simultaneous work on a lexical dictionary for distributed groups of people and enables cooperative work on the WordNet database. With WordVenture, the user can browse WordNet with a web application, and display its content in a graphical interface based on an interactive graph. It provides a user-friendly way for visualizing very large sets of contextual data. Displaying only selected nodes keeps the presentation clear. Functionality of traversing the graph by selecting nodes of interest allows to explore the semantic network. The user can also query WordVenture to find a specific word and display its senses and related concepts. Connections between nodes (words or senses) are illustrated as edges of a given type. To keep graphs clear, the user can set some constraints to visualize only required types of data [15].

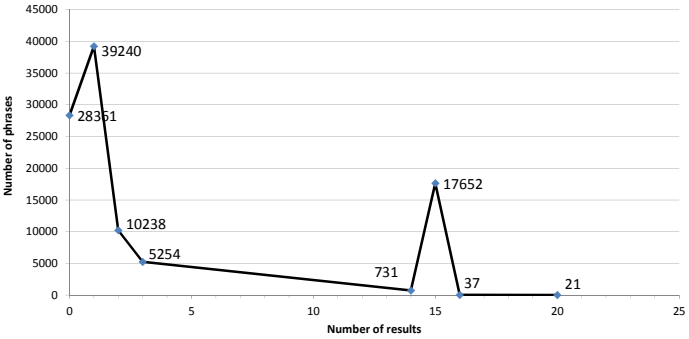

Fig. 1. Wikipedia pruning: results per query

V. WIKIPEDIA PRUNING

In the presence of a significant disproportion between the number of articles in Wikipedia and WordNet synsets, there is a need to pre-process Wikipedia and eliminate articles that are unlikely to be matched with WordNet synsets. The approach we took was to query Wikipedia via the Opensearch API with 117,798 words from WordNet. We set a limit to 20 results per query and found this way 340,000 matching articles.

We have prepared a series of statistics for the returned data. Figure 1 shows that almost half of the queries (43.87%) returned a unique result. The limit of 20 results per query has been reached only for 0.02% of queries, which indicates the parameter for the results limit is high enough.

In addition, 78,6% of articles is unambiguous (Table III), which compared to 51% of noun synsets defining only one phrase (Table II) is a rather high number. It is partially due to the fact that we recognize ambiguous phrases only if they occur both in WordNet and Wikipedia.

TABLE III
WIKIPEDIA PRUNING: PHRASES PER ARTICLES

Phrases	Articles
1	264959 (78,60%)
2	45156 (13,40%)
3	14324 (4,25%)
4	6076 (1,80%)
5	2839 (0,84%)
6	1529 (0,45%)
7 and more	2221 (0,66%)
Razem	337104

VI. MAPPING ALGORITHMS

Based on our analysis of WordNet and Wikipedia structure we have implemented algorithms, which automatically create mappings between these two databases. It is known that not all WordNet synsets can be mapped to Wikipedia articles. Often times general terms are not present in Wikipedia. For instance *friend* (a person you know well and regard with affection and trust) cannot be found in Wikipedia. The closest match we could find was *friendship*. However, more specific terms like *girlfriend* or *boyfriend* could be easily found. It is partially because WordNet is a dictionary whereas Wikipedia is an encyclopaedia. For the mappings to be useful we are less

 ISBN: 978-1-61208-206-6

interested in vague matches and we are looking for exact matches. We also prefer not to create a mapping than create a wrong one.

For that reason in our attempt we valued accuracy over coverage. The accuracy has been measured as a percent of correctly mapped synsets to all mapped synsets. The coverage is a percent of mapped synsets to all noun synsets. Note that mappings are many to many relations and sometimes we find more than one correct mapping for a synset or one article is related to more than one synset. A synset is considered to be correctly mapped when at least one of its mappings is correct.

The algorithm we constructed is combined from four independent approaches.

A. Unique results

The *unique results* algorithm was based on the fact that most of WordNet phrases are used in one synset only (Table II).

If a phrase is unique and querying Wikipedia returns only one result then we create a mapping. Such an observation allowed us to find related articles for 32,232 synsets (Table IV) which is 39% of all synsets. The evaluation for 100 random synsets has revealed an accuracy of 97% +- 3.34%. That gives us 32,024 mapped synsets out of 82,115 total synsets.

TABLE IV
UNIQUE RESULTS

Mappings	Articles
1	31987
2	118
3	3
Total	32232

The *xerox* synset is a good example where the algorithm works well.

> **Xerox, xerographic copier, Xerox machine** *[a duplicator (trade mark Xerox) that copies graphic matter by the action of light on an electrically charged photoconductive insulating surface in which the latent image is developed with a resinous powder]*

Searching for synonyms in Wikipedia gives following results.

> **Xerox**: 14 results
> 1. **Xerox** *[Xerox Corporation is an American multinational document management corporation that produced and sells a range of color and black-and-white printers, multifunction systems, photo copiers, digital production printing presses, and related consulting services and supplies.]*
> 2. **PARC (company)** *[(Palo Alto Research Center Incorporated), formerly Xerox PARC, is a research and co-development company in Palo Alto, California, with a distinguished reputation for its contributions to information technology and hardware systems.]*
> 3. **Xerox Star** *[The Star workstation, officially known as the Xerox 8010 Information System, was introduced by Xerox Corporation in 1981. It was the first commercial system to incorporate various technologies that today have become commonplace in personal computers, including a bitmapped display, a window-based graphical user interface, icons, folders, mouse, Ethernet networking, file servers, print servers and e-mail.]*
> ...
> **xerographic copier**: 0 results
> **Xerox machine**: 1 result
> 1. **Photocopier** *[A photocopier (also known as a copier or copy machine) is a machine that makes paper copies of documents and other visual images quickly and cheaply.]*

Applying the above described algorithm we create a mapping from the *Xerox* synset to the *Photocopier* article, which in fact have a redirect from *Xerox machine*.

It is easy to find an example where the algorithm does not work as expected. For instance the *indorsement* synset is matched with the *blank endorsement* article.

> **indorsement, endorsement, blurb** *[a promotional statement (as found on the dust jackets of books)]*
>
> **Blank endorsement** *[Blank endorsement of a financial instrument such as a check is only a signature, not indicating the payee.]*

It is because Wikipedia returns a single result for indorsement. The right *Testimonial* article is returned for the endorsement phrase, however, it is not matched as it is one of many.

> **Testimonial** *[In promotion and of advertising, a testimonial or show consists of a written or spoken statement, sometimes from a person figure, sometimes from a private citizen, extolling the virtue of some product.]*

B. Synonyms

In the presence of 21.4% synonyms in the pruned Wikipedia (Table III) and 49% in WordNet synsets (Table II), we assumed that if the same article occurs at least twice in the results from querying Wikipedia with synonym words from WordNet then a mapping exists. The *synonyms algorithm* has covered 22% of synsets with 88% +- 6.43% accuracy. That gives us 18,065 mapped synsets, 15,897 +- 1,161 of which are correct.

Harvard, Harvard University [a university in Massachusetts] is an example where the algorithm works well. Querying Wikipedia with the *Harvard* phrase gives us 14 results whereas *Harvard University* 13 results. Both queries return the *Harvard University* article at the top position in the result set, thus it is recognized as the correct one.

An example of a wrong mapping is for the **commission, delegacy, delegation, mission, deputation** *[a group of representatives or delegates]* synset. The algorithm creates an invalid mapping to the **Delegation** *[Delegation (or deputation) is the assignment of authority and responsibility to another person (normally from a manager to a subordinate) to carry*

 ISBN: 978-1-61208-206-6

out specific activities.] article, which is contained in results for *delegation* and *deputation*. The correct article **Delegate** *[A delegate is a person who speaks or acts on behalf of an organization (e.g., a government, a charity, an NGO, or a trade union) at a meeting or conference between organizations of the same level]* is to be found in the returned results, but it is further on the list.

C. Exact matches

A third implemented algorithm created a mapping whenever an article title and a synset phrase were the same, but only if the phrase was used in no more than one synset. As a result 59% of synsets have been matched with articles with a measured accuracy of 83% +- 7.35%. That gives us 48,447 synsets, 40,211 +- 3,560 of which are correct.

The strength of this algorithm lies in the fact that 51% of synsets have exactly one sense and define such unique terms as Lycopodium obscurum, Centaurea, Green Revolution, etc.

Among wrong results the **fishbone** *[a bone of a fish]* synset is to be found, which is mapped to the **Fishbone** *[Fishbone is a U.S. alternative rock band formed in 1979 in Los Angeles, California, which plays a fusion of ska, punk rock, funk, hard rock and soul.]* article. To our surprise manual search did not let us find any matching article.

D. Most used

The last approach was based on an assumption that the first returned result from the Wikipedia Opensearch API is the correct one. If a synset has synonyms, then we select an article that appears the most frequently and at the highest positions among all returned results. This trivial approach was tailored for improving the overall coverage. However, it has introduced a very high number of wrong mappings. As many as 84% synsets have been mapped with a measured accuracy of only 17% +- 7.36%. That gives us 68,976 mapped synsets, but with only 11,726 +- 5,047 correct.

E. Final results

The final run was selected based on the highest F-measure [16]. The F-measure is a weighted harmonic mean of precision and recall and it is defined with formula 1.

$$F = 2 * \frac{precision * recall}{precison + recall} \quad (1)$$

The precision is the number of correct results divided by the number of all returned results, whereas the recall is the number of correct results divided by the number of results that should have been returned. Mappings between synsets and articles can be correct, wrong or non-existing. To simplify calculations of the F-measure we assumed that all synsets can be mapped, thus the recall is the number of correctly mapped synsets divided by the sum of synsets, which are mapped correctly and not mapped at all.

It was an intersection of the Unique Results, Synonyms and Exact matches algorithms (2), which have produced the best results. The algorithms have been run separately and results

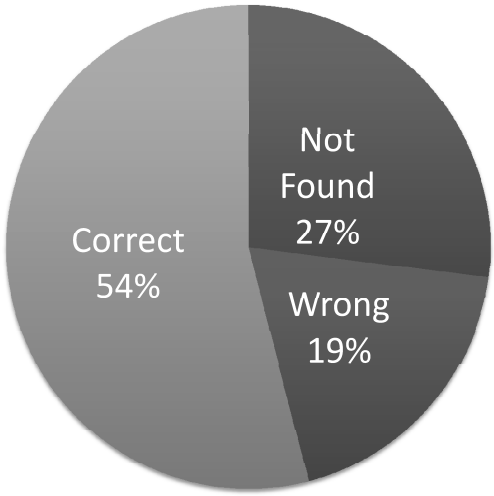

Fig. 2. Final results for Unique Results, Synonyms and Exact Matches

merged. In the effect 60,623 synsets were mapped, which is 74% of all noun synsets with a measured accuracy of 73% +- 8,7%, which is as many as 44,254 +- 5,247 correctly mapped synsets.

The overall results of running all four algorithms separately and in combinations are presented in Table V.

TABLE V
RESULTS OF MAPPING ALGORITHMS

Algorithm	Precision	Recall	F-measure
Unique results (UR)	0,97	0,38	0,55
Synonyms (S)	0,88	0,19	0,32
Exact matches (EM)	0,83	0,54	0,66
Most used (MU)	0,17	0,47	0,25
UR + S + MU	0,37	0,81	0,51
UR + S	0,86	0,43	0,58
UR + S + EM	0,73	0,68	0,70

VII. COLLABORATIVE EVALUATION AND CREATION OF MAPPINGS

Due to the nature of the problem, it is impossible to automatically evaluate created mappings to achieve higher precision. In order to speed up the process of evaluation and creation of missing mappings a system for collaborative work was implemented.

The project – ColabMap [17] enables many users to work simultaneously via the web interface. Their task is to assess correctness of automatically created mappings as well as to manually create new mappings.

The user needs to login in order to start assessing mappings. The authentication allows to track down already assessed items so that they are not presented to the same person twice, but but to resolve the problem of different opinions from different people. Next a random synset is displayed. If a mapping was created, an excerpt from a Wikipedia article is presented The user needs to choose one of four possible actions: Wrong, Acceptable, Perfect, or Skip. Skip should be chosen if the user

 ISBN: 978-1-61208-206-6

do not have enough expertise or certainty regarding accuracy of the mapping.

On the other hand, if a mapping does not exist yet or was wrong the user is asked to create a new mapping. In such a case the user is presented a list of possible articles from Wikipedia. There is also a field, which allows to search Wikipedia manually to find an article that is not on the list. The user may select multiple articles by choosing the Acceptable or Perfect score for each one, which will result in multiple mappings being created at once.

Answers of users are persisted separately so that if an administrator discovers a malicious user, they can be easily deleted. The results are presented on the statistics page. One can find there real-time statistics of evaluated and created mappings. There is also a feature, which allows to export mappings in a text format, but it is not yet exposed via the web interface.

The application back-end is written entirely in Java using the Spring framework. All data including WordNet and pruned Wikipedia are stored in the database. For the efficiency all Wikipedia queries and results are cached in the db as well. The module for accessing dictionaries and mappings can be easily decoupled from the web application and used in other applications through a well defined API. It allows to search for terms in both dictionaries making use of the established mappings.

The most current mappings between WordNet synsets and Wikipedia articles we deployed at web page of our ComputationalWikipedia project [18] aiming at create computational representations of Wikpedia [19].

VIII. Future work

Mappings between WordNet synsets and Wikipedia articles make it possible to use these two resources in Natural Language Processing simultaneously. We think the mappings should improve existing text representations used in the machine processing. The basic assumption is to provide extended information about words in the written text and using it provide elementary meaning of the utterances.

The integration of the resources opens possibilities to improve WordNet development. We plan to mine [20] Wikipedia structure and introduce new significant relations to WordNet. It should considerably extend the cross part of speech relations that are especially slimy defined in WordNet. We also plan to extend WordNet sparse synset definitions with extensive articles' content. Note that the definitions can be translated into other languages thanks to Wikipedia language links, which also enables multilingual linguistic dictionaries development.

ACKNOWLEDGMENTS: The work has been supported by the Polish Ministry of Science and Higher Education under research grant N N516 432338.

References

[1] http://royal.pingdom.com/2011/01/12/internet-2010-in numbers/, (Online, accessed: 2012 June).

[2] Y. Ding, D. Fensel, M. Klein, and B. Omelayenko, "The semantic web: yet another hip?" *Data & Knowledge Engineering*, vol. 41, no. 2-3, pp. 205–227, 2002.

[3] K. Goczyła, T. Grabowska, W. Waloszek, and M. Zawadzki, "The knowledge cartography–a new approach to reasoning over description logics ontologies," *SOFSEM 2006: Theory and Practice of Computer Science*, pp. 293–302, 2006.

[4] P. Buitelaar, P. Cimiano, and B. Magnini, "Ontology learning from text: An overview," *Ontology learning from text: Methods, evaluation and applications*, vol. 123, pp. 3–12, 2005.

[5] C. Manning, H. Schütze, and MITCogNet, *Foundations of statistical natural language processing*. MIT Press, 1999, vol. 59.

[6] G. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[7] J. Szymański, *WordNet - bazodanowy system jako slownik jezyka angielskiego*. Politechnika Gdańska, 2006.

[8] http://www.google.com/adplanner/static/top1000, (Online, accessed: 2012 June).

[9] Wikipedia Statistics, http://stats.wikimedia.org/EN, (Online, accessed: 2012 June).

[10] M. Ruiz-Casado, E. Alfonseca, and P. Castells, "Automatic assignment of wikipedia encyclopedic entries to wordnet synsets," *Advances in Web Intelligence*, pp. 380–386, 2005.

[11] J. Szymański and D. Kilanowski, "Wikipedia and WordNet integration based on words co-occurrences," *Proceedings of 30th International Conference Information Systems, Architecture and Technology*, vol. 1, pp. 93–103, 2009.

[12] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: A large ontology from wikipedia and wordnet," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 3, pp. 203–217, 2008.

[13] F. Suchanek and G. Kasneci, "Yago: a core of semantic knowledge," *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, 2007.

[14] Szymański, J. et al. (2012, June) Cooperative Wordnet Editor – WordVenture. http://wordventure.eti.pg.gda.pl.

[15] J. Szymański, "Cooperative wordnet editor for lexical semantic acquisition," in *Knowledge Discovery, Knowlege Engineering and Knowledge Management*, ser. Communications in Computer and Information Science, A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe, Eds. Springer Berlin Heidelberg, 2011, vol. 128, pp. 187–196.

[16] G. Hripcsak and A. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[17] J. Szymański and R. Korytkowski. (2012, June) ColabMap project. http://kask.eti.pg.gda.pl/colabmap.

[18] Szymański, J. et al. (2012, June) Computational Wikipedia project. http://kask.eti.pg.gda.pl/CompWiki/index.php?page=wordnet&.

[19] J. Szymański and W. Duch, "Representation of hypertext documents based on terms, links and text compressibility," *Neural Information Processing. Theory and Algorithms*, pp. 282–289, 2010.

[20] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

 ISBN: 978-1-61208-206-6