Automated Extraction of Lexical Meanings from Corpus: A Case Study of Potentialities and Limitations

Maciej Piasecki[1]

Instytut of Informatics, Politechnika Wroclaw University of Technology
maciej.piasecki@pwr.wroc.pl, www.plwordnet.pwr.wroc.pl

Abstract. Large corpora are often consulted by linguists as a knowledge source with respect to lexicon, morphology or syntax. However, there are also several methods of automated extraction of semantic properties of language units from corpora. In the paper we focus on emerging potentialities of these methods, as well as on their identified limitations. Evidence that can be collected from corpora is confronted with the existing models of formalised description of lexical meanings. Two basic paradigms of lexical semantics extraction are briefly described. Their properties are analysed on the basis of several experiments performed on Polish corpora. Several potential applications of the methods, including a system supporting expansion of a Polish wordnet, are discussed. Finally, perspectives on the potential further development are discussed.

1 Introduction

A technique or tool must have its purpose and justification for using it Moreover, before we apply it, we should ask how it works and whether it does what was promised. We are going to approach these questions with respect to the automated extraction of formalised descriptions of lexical meanings from corpora.

Automatic methods are based on algorithms. An algorithm requires a formal model of the processed data: input and output. The crucial issue is a formal, or at least formalised, description of lexical meanings which will be discussed in Section 2. Having a formal model of lexical meanings defined, next we need to analyse what kind of evidence pertaining to the elements of the model can be automatically extracted from a corpus. A short review of the possible sources of evidence is presented in Section 3. Two basic paradigms of extraction of lexical meanings, namely a *pattern-based* and *distribution-based*[1] are introduced in Sections 4 and 5, respectively. Next, various applications of the introduced techniques and their hybrid combinations in the area of Computational Linguistics, but also Linguistics in general, are discussed in Section 6. In spite of significant improvements observed in the field during the last decade still in many aspects we should talk rather about potentialities instead of fully developed solutions. Perspectives on the possible development of the methods in the nearest future is briefly presented in Section 7 which concludes the paper.

2 Formalised Description of Lexical Meanings

From the historic perspectives, the oldest form of lexical meaning description is a lexical entry based on listing different *lexical units* (meanings) of a *lemma* in separate positions. Each lexical unit is given a short description in the natural language and often accompanied by a set of examples. This kind of lexical meaning definition is not very useful as an output for the extraction algorithms, since it is not formalised and a synthesis of a proper definition expressed in the natural language is a serious demand for Natural Language Processing. Proper lexicon definitions, e.g. in a sense postulated by [1], occur very rarely in a general corpus, unless it includes a dictionary. Automatic construction of precise definition would require detailed model of the structure and meaning of natural language definitions, i.e. utterances of special kind and purpose.

[1] The latter one is also called clustering-based [20].

On the opposite end of formalisation scale lays a technique based on the application of *meaning postulates* to the semantic description of logical symbols representing lexical units. Precise semantic representation of the natural language is based on a formal, logical language, in which lexical meanings are represented by logical predicates. As the predicate meaning is defined by its formal interpretation the only way to transfer meanings from natural language units to predicates is by constraining interpretation of the predicates, i.e. by shaping the formal structures of the interpretation model. A meaning postulate is an axiom constraining the possible interpretation of a given logical predicate used for representing meaning of a particular lexical unit, e.g. [8]

Defining a meaning postulate is usually a demanding task, which requires rich knowledge. The tasks complicates with the increasing size of the set of meaning postulates. One can hardly imagine the acquisition of meaning postulates from text corpora, as it would require extraction of detailed, formally expressed, knowledge about the world.

In Componential Semantics lexical meaning is defined as a set of features which distinguish it from other lexical meanings. The meaning can be also presented as a expression in a formalised language. The expression consists of basic meaning atoms linked by some operators, e.g. a short review of works in this area given by Dowty [7] but also works of Wierzbicka [35] or Pustejovsky [29] or Mel'čuk [17] can be mentioned here as examples of component-based analyses.

In traditional thesauruses, e.g. Roget's Thesaurus, lexical semantic relations like synonymy, hypernymy, meronymy etc. and grouping lexical units into semantic clusters are utilised in constructing large coverage descriptions of lexical meanings. Description by a network of relations defined on the set of lexical units is partial in comparison to the componential analysis, however appeared to be useful not only for human readers but also for Language Technology applications, e.g. hundreds of applications of WordNet – an electronic thesaurus of English [9, 26]. Lexical semantics relations occur also as elements of formalised lexical meaning descriptions, e.g. [18, 29].

In sum, there are two basic types of elements constituting formal descriptions: basic components and relation instances. However, the components must later be combined into complex expressions to form definitions. This complicates the process of automated extraction. That is why we will focus mainly on automatic extraction of lexical semantic relations.

3 Machine Tractable Evidence

Analysing a corpus we can take two possible perspectives. First we can concentrate on detailed analysis and drawing conclusions from particular occurrences of the given lemma in focus, trying to extract detailed information concerning its semantics from each context of occurrence. Secondly, we can take a global perspective and analyse all occurrence with lower accuracy each but taking into account statistical evidence. Both approaches are used and will be discussed in the following subsections.

3.1 Embedded definitions

Since an utterance of a text performs often informative function, we can find language expressions in a corpus which describe meanings of particular lemmas in various ways, e.g. an example of quasi-definition found in corpus by Hearst [13]

> The *bow lute*, such as the *Bambara ndang*, is plucked and has an individual curved neck for each string.

The above sentence relates *Bambara ndang* to the *bow lute* as its hyponym, but also characterises it by some details. In many cases, we can identify specific lexico-syntactic constructions that indicate a pair of lemma as an instance of a particular lexical semantic relation, e.g. hypernymy in the above example.

When we take into consideration more complex lexico-syntactic and possibly also semantic structures, we can also try to identify complex, descriptive definitions included in a text, e.g. the descriptive, ending part of the example above, or an example given in [10, pp. 3]:

> A linguist is a scientist who investigates human language [...]

Such indicative language constructions are more likely to be found in texts of specific genres like encyclopaedias (almost every entry includes a helpful passage of text) or text books, but even experiments performed on a general corpus bring relatively good results, see Section 4.

In order to utilise the information expressed in this way, we need to apply some kind of structural analysis which is sensitive to the lexico-syntactic structures in focus. This issue will be discussed in details in Section 4.

3.2 Distributional Hypothesis

As the structure of language expressions is determined by the properties of constituents, including semantic properties, an analysis of a large number of language expression occurrences should allow us to identify some regularities of semantic nature. This general assumption has appeared in several linguistic theories. A well known version was proposed by Harris [11] in the form of Distributional Hypothesis.

Harris [11] in his *Distributional Hypothesis* expressed a strong belief that there is a direct relation between the observed use of language expressions and their meaning (cited after to [32]):

> The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities.

The granulation of entities is not specified in the hypothesis. They can be any language items. Henceforth, we will concentrate our attention on one or multiple word *lemmas*, representing one or several *lexical units* each. A lemma is expressed in text by one of several possible *word forms*.

As in the whole corpus it is hard to identify strict restrictions on lemma combinations that are always preserved, it is more practical to perceive restrictions as preferences of different strength. It is used to describe a set of restriction imposed on a lemma in a kind of reversed way by identifying a set of particular *contexts* in which the given lemma can occur. However, we should describe each context by the strength of preference too.

A context is a textual situation involving a number of word form occurrences. A more detailed definition requires answering two questions:

1. what kind of features do we use to describe a context,
2. and how large is a context.

As for the first question, an occurrence context of the lemma `x` can be identify with:

- the association of `x` with a particular textual object, e.g. a particular document `d`,
- the co-occurrence of `x` with a particular word form or lemma `y`,
- the participation of `x` in an instance of a particular lexico-syntactic relation, e.g. an occurrence of `x` as an argument of a particular predicate together with some or all argument set, see detailed examples in Section 5.

The size of the context can be defined as the size of the whole textual object, a *text window* (i.e. a text snippet of certain number of words) or some syntactic construction like a sentence.

Identification of contexts in which a given lemma `x` can occur and the evaluation of the association strength result in a model of the distribution of `x`. According to the Distributional Hypothesis, comparison of the distributions of two lemmas allows us to draw conclusions concerning the similarity of their meanings. This claim is verified on the basis of empirical data in Section 5.

4 Pattern-based Relation Extraction

Traditional semantic lexicons consist of entries written in the natural language, but all definitions have similar structure and associate described lemmas with other lemmas occurring in entries in ways indicating certain lexical semantic relations, e.g. hypernymy (hypernyms are usually given at

the beginning of a lexicon entry description). Following this observation, a number of approaches to the extraction of lexical semantic relations from Machine Readable Dictionaries were proposed. A set of lexico-syntactic patterns was defined, where each pattern was a regular expression[2] defined over word forms, their morpho-syntactic properties and/or simple syntactic structures.

Text in a large corpus had seemed to be of much less predictable character, however, as the seminal work of Hearst [13] showed, similar patterns can be applied to the corpus text and produce valuable results. One of the productive Hearst's pattern is presented below:

NP_0 ..... such as {NP_1, NP_2 . . . . (and | or)} NP_n

The pattern or more precisely the language constructions identified by the pattern in corpus implies that each noun phrase NP_i is a hyponym of the noun phrase NP_0, i.e. the hypernymy relation holds between lemmas represented in the text by the given noun phrases. Hearst [12, 13] constructed manually only five patterns frequently matched in a corpus and appealingly accurate. The accuracy was measured as a number of lemma pairs linked by the hyponymy relation in WordNet [9] to all those extracted. For the pattern shown above, for example, 61 of 106 extracted lemma pairs from Grolier Encyclopedia were confirmed in WordNet [13].

The implicit assumption here is that one can construct patterns accurate enough to draw correct conclusions from single occurrences of lemma pairs. In general, however, it seems barely possible due, amongst others, to the presence of metaphor. Without deeper semantic and pragmatic analysis, instances of metaphor may be hard to distinguish from literal uses. Hearst extracted *aeroplane* as a hyponym of *target* and *Washington* as an instance of *nationalist*; such derived associations are clearly specific to particular documents from which they were extracted. Another problem is the scarcity of pattern instances in corpora; merely 46 instances were acquired from 20 million words of the New York Times corpus [13].

These patterns are expressed in a grammar of limited expressive power and work on the basis of an assumption of the fixed linear order of English sentence. For a highly inflected language, like Polish, a more sophisticated mechanism is required. Thus we applied a language of morpho-syntactic constraints called JOSKIPI utilised in the TaKIPI tagger of Polish [22]. JOSKIPI is equipped with operators for testing morpho-syntactic properties of particular words, their compatibility (including agreement), defining sequences and iterating tests over word groups of non-predefined size (e.g. until the beginning of a sentence or the fulfilment of a condition). Six productive patterns were defined, cf [26], and a scheme of one of them is presented below[3]:

```
NP1 (Adj|Adv|Noun|,)*
(base2{i, oraz}(and)) (base2{inny, pozostały}(other, remaining), nmb=pl)
(Adj|Adv)* NP2(cas=cas(NP1))
```

The pattern identifies the lemma (potentially a multiword one) of NP1 as a hyponym of the lemma of NP2. Two other similar patterns were constructed on the basis of such lexical markers like: *taki jak* (*such as*) and *w tym* (*including*). An example of the construction covered by the pattern including *taki jak* is presented below:

Betondour doskonale nadaje się do wykończania podłóg w pomieszczeniach takich jak garaże,
Betondour perfectly is suitable for dressing floors in rooms like garages,
warsztaty samochodowe, magazyny, sklepy, pomieszczenia produkcyjne, piwnice czy wykonane
garages, stores, shops, manufacture rooms, cellars or made
z betonu schody.
from concrete stairs.

As all three types of language construction are used in a very similar role, these three patterns were merged together as three variants and tested jointly. We applied them to extract hypernymic pairs from three corpora:

– *IPI PAN Corpus* (including about 254 million tokens) [28],

[2] It has the expressive power of a regular grammar.

[3] In a simplified form, the original JOSKIPI expressions have been exchanged to labels describing words and word groups identified.

– a corpus of the electronic edition of a Polish newspaper *Rzeczpospolita* from January 1993 to March 2002 (about 113 million tokens) [31];
– and a corpus of large texts in Polish (about 214 million tokens) collected from the Internet; only documents containing a small percentage of erroneous word forms (tested manually) and not duplicated in the other two corpora were included in the collected corpus.

Henceforth, we will refer to all three corpora used together as the *joint corpus*. In order to increase precision, we limited the application of the patterns only to cases in which two nominal lemmas from the predefined list occurred in the same sentence. 13 285 nominal lemmas have been collected from: the core part of plWordNet [5, 6] – 5340, a small Polish-English dictionary [27], two-word lemmas from a general dictionary of Polish [30], and the IPI PAN Corpus [28] – only those that occur over 1000 times.

The results are presented in Table 1. The accuracy was manually measured on the basis of a representative sample of the produced pair list as a ratio of positively assessed pairs to all extracted. Each pair linked by a hypernymy relation – possibly not directly, with any number of intervening other lemmas – was counted as a positive case while others as negative ones.

IPI PAN Corpus		Corpus from the Web		*Rzeczypospolita* Corpus		the joint corpus	
No. of pairs	Accuracy	No. of pairs	Accuracy	No. of pairs	Accuracy	No. of pairs	Accuracy
14611	30.06%	5983	32.52%	6682	33.16%	24437	30.69%

Table 1. The results of hypernymy extraction by manually constructed lexico-morphosyntactic patterns.

The extracted list cannot be used directly as a list of hypernymic pairs – the accuracy is too low (however it matches the level achieved by other approaches). The accuracy should be increased above 50% to make the tool interesting for linguists. When this merged pattern is combined with two other ones, the accuracy can be increased up to 41.05% for the price of the reduced number of pairs to 8777 [26].

The pattern-based approaches most often target the hypernymy relation. However, manually constructed patterns were also applied to the extraction of meronymy, too, e.g. [2].

Manually constructed patterns achieve relatively high precision for the cost of the limited number of pairs extracted and some time spent on their manual tuning on the basis of corpus analysis. Due to their expressive power they are difficult to be extracted automatically. However, Pantel and Pennacchiotti [20] with their *Espresso* algorithm and recently Kurc and Piasecki with its modified version and adapted to Polish [15], called *Estratto*, showed that a set of simpler, general patterns can be effectively extracted and applied in a way resulting in accuracy even better on huge corpus than the manually build patterns. *Espresso*/*Estratto* can be applied to any lexical semantic relation that is manifested in corpus by some lexico-syntactic markers, e.g. Espresso was successfully used to extract hypernymy but also other types of relations like e.g. *part-of*, *reaction* (in chemical sense) or *production*. Both algorithms work according to the same schema:

1. A set of example instances of the relation in focus – pairs of lemma associated by the relation, is delivered to the algorithm.
2. All close co-occurrences of lemmas from the same pair (e.g. in the same sentence) are identified in the corpus and patterns are generated in a form of generalised descriptions of token sequences occurring in between the pairs of lemmas.
3. A measure of reliability is calculated for each pattern on the basis of instances covered by it and their reliability (the reliability of example instances is set to 1).
4. A subset of the highest ranked patterns is stored and next used to extract new instances.
5. Finally, the reliability of the extracted instances is calculated in similar way on the basis of the patterns matching the instances in corpus and their reliability; only the highest ranked instances are kept for the next iteration.

Correct hypernymy instances	
koncesja (*concession*)	decyzja (*decision*)
kapłan (*priest*)	człowiek (*human*)
maj (*May*)	okres (*period*)
kwestia (*issue*)	problem (*problem*)
sowa (*owl*)	ptak (*bird*)
klient (*customer*)	osoba (*person*)
pielęgniarka (*nurse*)	osoba (*person*)
profesor (*profesor*)	człowiek (*human*)
galeria (*gallery*)	miejsce (*place*)
matematyka (*mathematics*)	przedmiot (*subject*)
matka (*mother*)	kobieta (*woman*)
helikopter (*helicopter*)	maszyna (*machine*)
droga (*way*)	szlak (*track*)
zespół (*team*)	grupa (*group*)
mecz (*game*)	spotkanie (*meeting*)
restrukturyzacja (*restructurisation*)	zmiana (*change*)
konsument (*consumer*)	osoba (*person*)
tenis (*tennis*)	sport (*sport*)
festiwal (*festival*)	impreza (*event*)
dziennik (*daily*)	dokument (*document*)
medycyna (*medicine*)	nauka (*science*)
anioł (*angle*)	istota (*being*)
spółka (*partnership*)	firma (*firm*)
szczur (*rat*)	szkodnik (*pest*)
skorpion (*scorpio*)	znak (*sign*)
rak (*cancer*)	choroba (*illness*)
nagroda (*prize*)	wyróżnienie (*distinction*)
Non-hypernymy associations	
przepis (*recipe*)	kwestia (*issue*)
silnik (*engine*)	jednostka (*unit*)
człowiek (*human*)	drzewo (*tree*)
program (*program*)	działanie (*activity*)
muzyka (*music*)	dźwięk (*sound*)
istota (*being*)	nic (*nothing*)
wojsko (*army*)	organizacja (*organisation*)
stowarzyszenie (*association*)	instytucja (*institution*)
cień (*shadow*)	wróg (*enemy*)
książka (*book*)	materiał (*material*)
słońce (*sun*)	czynnik (*factor*)

Fig. 1. Examples of lemma pairs extracted from the joint corpus by the application of the merged group of patterns including the *i inny* (*and other/remaining*) pattern.

The application of *Estratto* initiated by a list of hypernymic pairs from plWordNet to the IPI PAN Corpus [15] produced 25 361 pairs with the accuracy of 41% (measured manually on a representative sample in the same way as for the manual patterns). The result obtained automatically is significantly better than the one produced by the manual patterns. Preliminary results of the application of Estratto to the extraction of meronymy and adjectival antonymy are promising, in spite of the significantly worse accuracy on the level around 30%.

A weak point of the pattern-based approaches is that each occurrence of a lemma pair matching the pattern results in extracting it. There are many accidental associations. However, this occurrence sensitivity can be also an advantage for a linguist, as we can easily trace back from an extracted pair to the place of its occurrence in corpus, e.g. one can list all pairs not supported by a thesaurus. In the case of pairs extracted by automatically created patterns, the situation is slightly different, as an extracted instance is mostly supported by more than one general pattern.

5 Distributional Semantics

According to the interpretation of the Distributional Hypothesis presented in Section 3.2 comparison of distribution models of particular lemmas can result in some assessment of 'how close' meanings of both lemmas are. It is important to emphasise that in the case of most distributional methods model of the distribution is built jointly for the whole lemma and the influence of its different lexical units is mingled in it.

The basic result of distributional methods is a *Measure of Semantic Relatedness* (MSR). MSR is a function which for a lemma pair returns a number expressing the strength the semantic relation between them, i.e. $\mathsf{MSR} : \mathsf{L} \quad \mathsf{L} ! \quad \mathsf{R}$, where L is a set of lemma and R is a set of real numbers.

Many methods have been proposed for MSR extraction, but they all contain four general steps, more or less clearly delineated.

1. *Corpus preprocessing* – typically up to the level of shallow syntactic analysis.
2. *Co-incidence matrix construction* – in which rows correspond to lemmas being described and columns to contexts; each cell $\mathsf{M}[\mathsf{x}_i ; \mathsf{c}_j]$ stores the frequency of the occurrences of the lemma x_i in the context c_j.
3. *Matrix transformation* – a possible reduction of size and/or combination of feature *weighting* and *selection*.
4. *Semantic relatedness calculation* – lemma descriptions are compared by the application of an assumed measure of similarity between row vectors.

Depending on the type of the context used, [19] distinguishes between measures of *semantic relatedness* and *semantic similarity*. Semantic relatedness is obtained on the basis of contexts defined as co-occurrence with a particular lemma in one document or a text window, i.e. the types: 1 and 2 on the page 3. Semantic similarity is extracted on the basis of lexico-syntactic relations used as contexts – the type 3, e.g.
`X` occurs as `subject_of(`*`a particular verb`*`)` or as `modified_by(`*`a particular adjective`*`)`.
According to our experiments performed on the IPI PAN Corpus, cf [23], semantic relatedness encompasses broader semantic associations among lemmas, based on co-occurrences of both lemmas in the description of the same situation. According to [19] lemma pairs receiving high values of semantic similarity should represent lexical-semantics relation used in thesauruses, e.g., synonymy, hypernymy, meronymy, etc. However, intermediate methods are quite conceivable. For example, one can combine lexico-syntactic constraints with co-occurrences in the description of context. So, there is a continuum of methods with these two extremes. Semantic relatedness is a more general notion as among lemma pairs expressing high semantic relatedness one can also find lemma pairs expressing high semantic similarity.

An example of a list of 20 top semantically related lemmas to the given one is presented in Figure 2. The list was produced by a MSR extracted from the joint corpus and for l3 285 nominal, both discussed in Section 4. The MSR was based on the following types of lexico-syntactic contexts:

1. modification by *a specific adjective* or *a specific adjectival participle* (41 619 features),

2. co-ordination with a a specific noun (115 604),
3. modification by a specific noun in the genitive case (115 604),
4. occurrence of a specific verb for which a given noun lemma can be its subject (19 665),

There were 167 834 active features left after weighting and selecting.

In Figure 2 we can notice that the list includes hypernyms of *gaz ziemny* (*natural gas*), e.g. *gaz* (*gas*) and kopalina (*mineral, resource, fossil*); co-hyponyms, e.g. *węgiel kamienny* (*coal (pit-coal)*) and *ropa* (*oil*); loosely related cousins from the broader part of the hypernymic structure, e.g. *azot* (*nitrogen*) and *cynk* (*zinc*) and also lemmas similar because of the similarity of the use of the respective substances e.g. *biokomponent* (*biocomponent*) (as an addition to car fuel), while *gaz ziemny* can be used as the car fuel by itself.

gaz ziemny (*natural gas*)	
gaz (*gas*)	0.258
węgiel kamienny (*coal (pit-coal)*)	0.207
węgiel brunatny (*brown coal*)	0.197
ropa (*oil*)	0.193
olej opałowy (*heating oil*)	0.164
paliwo (*fuel*)	0.161
wodór (*hydrogen*)	0.160
kopalina (*mineral, resource, fossil*)	0.160
węgiel (*coal*)	0.143
olej napędowy (*diesel fuel*)	0.140
gaz płynny (*liquid gas*)	0.140
koks (*cox*)	0.127
ołów (*lead*)	0.119
azot (*nitrogen*)	0.119
tlen (*oxygen*)	0.116
uran (*uranium*)	0.116
biokomponent (*biocomponent*)	0.115
cynk (*zinc*)	0.114
łupek palny (*slate (fuel)*)	0.113
benzyna (*gasoline*)	0.110

Table 2. A list of the 20 lemmas most similar to the given one according to the MSR from [26].

There is a notorious problem with the evaluation of MSRs. Manual inspection is misleading – one can always find good and bad examples on the lists of the most related lemmas. Simple calculation of a kind of accuracy on the basis of lemma pairs occurring on the lists and representing particular lexical semantic relations tells only part of the truth. MSR generates a function assigning some strength of semantic relatedness to every lemma pair. The assignment of higher values to pairs representing some relations is very expected but it does not exhaust the potential of MSR. Manual assessment of the MSR values is not feasible and there are no similar manually created resources to compare with. Among several potential methods discussed e.g. in [24, 36], application of MSR as the only knowledge source in solving a synonymy test, which was originally conceived for humans, have been fruitfully applied in several experiments.

Besides MSR, another kind of lexical semantic knowledge extracted on the basis of statistical evidence are semantic selectional restrictions of predicates [16]. For each subcategorisation frame of a lemma and for each argument position of the frame semantic classes of language expressions occurring on the given position are identified. As extraction of the selectional restrictions is based on statistical evidence, the association of classes to frame argument positions is described by the strength of association. The main problems are: definition of semantic classes, recognition of the occurrences of frames and classes in text and, finally, identification of statistically significant associations: a position – a class. A set of classes is defined in relation to some semantic lexicon

(e.g. semantic codes assigned to lexical units) or a thesaurus. In the latter case, the hypernymy structure (of lexical units or semantic field) is used and a subset of the nodes of the structure is selected as a basis for the classes, i.e. a class corresponds to a node. Frame occurrences are identified in text by some means of parsing (mostly shallow parsing) and class occurrences by some form of Word Sense Disambiguation (as one lemma can represent several classes). Next, the frequencies of a particular frame argument position – a particular class co-occurrences are collected and a kind of transformation based on weighting and/or selection is applied in a way similar to the transformations described for MSR.

6 Applications

Accuracy of any single method of extraction is around 30% in comparison to manually constructed lexical semantic resources, e.g. to relations described in plWordNet – a Polish wordnet [26]. However, different methods extract lexical semantic relations of different types, e.g. a kind of semantic relatedness (MSRs) vs a particular relation extracted by the given lexico-syntactic pattern. The methods differ also in the character of the extracted data: continuous space of relatedness values vs discrete, binary information about lemma pairs. Finally, the methods differ also in types errors made. Thus, a combination of several methods, i.e. several sources of evidence should provide a broader perspective and more accurate description (for the latter it is important that the errors are not made in similar ways by the methods). The first example of successful combination is the system *Estratto* [15] discussed earlier. Estratto combines the application of patterns with their evaluation based on statistical evidence. As a result, the output is not only a list of lemma pairs but each pair is described by its reliability value based on statistical evaluation.

Another example can be the WordNet Weaver system supporting linguists in extending the nominal part of the plWordNet thesaurus [26]. For each new lemma (i.e. not described in plWordNet yet), WordNet Weaver generates a set of subsets of the hypernymy structure (i.e. subgraphs comprising several linked lexical units each) as attachment suggestions. A suggestion expresses a possible area in the hypernymy structure in which one of the lexical units of the given new lemma should be added as a hypernym, hyponymy or a synonym. Suggestions are based on the combined evidence coming from 4 types of knowledge sources. The sources were all extracted for 13 285 Polish nominal lemmas from the joint corpus discussed in Section 4 and are characterised below:

- a high accuracy MSR based on the Rank Weight Function, see [26],
- post-filtering lemma pairs produced by the selected MSR with a classifier presented in [25] – the percentage of lexico-semantic relation instances increases among the filtered pairs
- manually constructed lexico-morphosyntactic patterns described in [26], including the pattern presented in Section 4,
- and the results generated by the *Estratto* algorithm application [14, 15] discussed in Section 4.

All knowledge sources, except *Estratto*, were extracted from the joint corpus discussed earlier. Estratto was applied only to the IPI PAN Corpus and the corpus of *Rzeczpospolita*.

WordNet Weaver has been successfully applied in expanding plWordNet by 15 200 new lexical units (8 700 new synsets). The manual workload was 3.4 person-months. We observed significant improvement in the pace of work in comparison to a purely manual work. Detailed evaluation of WordNet Weaver on the basis of the analysis of the work of linguists, as well on the basis of automated tests can be found in [26]. Here, we would like to highlight only that in the case of 75.24% of new lemmas at least one genereted suggestion was found to be helpful by the linguist.

WordNet Weaver is an example of the application of the extraction methods as a support for the construction of thesauruses. The automated methods can also serve as critiques of the existing thesauruses and lexicons facilitating discovery of lacking informations or identifying potential errors (cf [26] for the experience of this kind collected with WordNet Weaver).

A MSR can be used to generate clusters of lemmas associated semantically in the domain represented by the corpus. The automated methods can facilitate analysis and comparison of systems of lexical meanings characteristic for particular domains represented by domain corpora.

The automated methods supports naturally a data-driven approach in which data collected in a large corpus are the primary source. Thus, they deliver a perspective which is objective to the extent of the corpus dependence, but not influenced by the subjective interpretation of a researcher. In [21], we presented a comparison of the lemma associations obtained from surveys of human informants vs the associations extracted from the IPI PAN Corpus with the help of a MSR. The intersection of both lists is around 20% only, but the automatically extracted data presents also a valuable perspective on the understanding of Polish *collective symbols* and *flag words*.

7 Perspectives

In spite of more than 20 years of history of Distributional Semantics methods, there is still room for further development. First, still larger and larger corpora are available for many languages lacking such corpora earlier. Our experience with applying different types of contexts, e.g. [23, 24], shows that one can expect better accuracy of MSRs based on a deeper lexico-syntactic analysis. In the case of Polish, we have not had any robust parser at our disposal yet. Knowing the limitations of the lexico-morpho-syntactic constrains applied, we tried to increase their accuracy for the price of the unavoidable decrease in their recall. However, in this way only a portion of information included in text has been utilised. There are several potential factors influencing negatively the accuracy of MSR but probably the most important one is the fact that MSR is constructed for lemmas on the basis of evidence collected from lemma occurrences. As a result, in the majority of cases one sense, or most two senses (lexical units), dominate in the upper part of the list of the most semantically related lemmas generated for a given lemma. However, a potential improvement would require previous disambiguation of the corpus with respect to word senses (a task which is more difficult than the MSR construction) or a new method of MSR construction in which both processes, i.e. meaning extraction and word sense delimitation would be performed in parallel.

Combined methods seem to be in their initial stage of development, especially multi-criteria methods of automated resource construction, like WordNet Weaver [26] or [33]. We should see an intensive development of this subfield.

The extraction methods are gradually becoming more accessible to users not familiar with the technical details of the Language Technology, e.g. to linguists, as there are various attempts to make the language tools available and accessible to such users, e.g. the Clarin project[4]. Among the goals of Polish part of Clarin is to make the SuperMatrix system [3] available as a part of this research infrastructure. Among other options, SuperMatrix delivers tools for the pattern-based approach and Distributional Semantics. It was used for the generation of the majority of examples presented in this paper. Accessibility of the technology should boost the cooperation between linguists and researchers working in the area discussed here. Such a cooperation is needed.

Measure of Semantic Relatedness can be perceived as defining a specific perspective on lexical semantics.

An important future area of development for the extraction methods are multilingual applications i.e. extraction of lexical semantic relations and MSR from multilingual corpora in a way synchronised or mutually related. Potential results should be valuable for the construction of multilingual resources for lexical semantics and applications of the Language Technology, as well as possible linguistic comparative studies.

[4] `www.clarin.eu`

References

[1] J. D. Apresjan. *Semantyka leksykalna. Synonimiczne środki języka [Lexical semantics].* Warszawa, 2000. translated by Z. Kozłowska and A. Markowski.

[2] Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 57–64, Morristown, NJ, USA, 1999. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1034678.1034697.

[3] Bartosz Broda and Maciej Piasecki. SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. In Mieczysław A. Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, Advances in Soft Computing, pages 345–352, Warsaw, 2008. Academic Publishing House EXIT.

[4] Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006. The Association for Computer Linguistics.

[5] Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. Words, Concepts and Relations in the Construction of Polish WordNet. In A. Tanâcs, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, editors, *Proceedings of the Global WordNet Conference, Seged, Hungary January 22–25 2008*, pages 162–177. University of Szeged, 2008.

[6] Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, and Magdalena Zawisławska. plWordNet — The Polish Wordnet. URL `www.plwordnet.pwr.wroc.pl`. Online access to the database of plWordNet: `www.plwordnet.pwr.wroc.pl`, 2009.

[7] David R. Dowty. *Word Meaning and Montague Grammar*, volume 7 of *Synthese Language Library*. D. Reidel Publishing Company, Dordrecht:Holland/Boston:U.S.A./London:England, 1979.

[8] David R. Dowty, Robert E. Wall, and Stanley Peters. *Introduction to Montague Semantics*, volume 11 of *Synthese Language Library*. D. Reidel Publishing Company, Dordrecht:Holland/Boston:U.S.A./London:England, 1981.

[9] Christiane Fellbaum, editor. *WordNet – An Electronic Lexical Database*. The MIT Press, 1998.

[10] Victoria Fromkin, Bruce Hayes, Susan Curtiss, Anna Szabolcsi, Tim Stowell, Edward Stabler, Dominique Sportiche, Hilda Koopman, Patricia A. Keating, Pamela Munro, Nina Hyams, and Donca Steriade. *Linguistics: An Introduction to Linguistic Theory*. Blackwell Publishing, 2000.

[11] Zellig Sabbetai Harris. *Mathematical Structures of Language*. Interscience Publishers, New York, 1968.

[12] Marti A. Hearst. *Automated Discovery of WordNet Relations*, chapter 5, pages 131–151. Volume 1 of , Fellbaum [9], 1998.

[13] Matti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceeedings of COLING-92*, pages 539–545, Nantes, France, 1992. The Association for Computer Linguistics.

[14] Roman Kurc. Automatyczne wydobywanie leksykalnych relacji semantycznych na podstawie prostych wzorców syntaktyczno-leksykalnych. Master's thesis, Faculty of Computer Science and Management, Wroclaw University of Technology, 2008.

[15] Roman Kurc and Maciej Piasecki. Automatic Acquisition of Wordnet Relations by the Morpho-Syntactic Patterns Extracted from the Corpora in Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 3nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)*, pages 181–188, 2008.

[16] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 2001.
[17] I. A. Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.
[18] I.A. Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988.
[19] Saif Mohammad and Graeme Hirst. Distributional measures as proxies for semantic relatedness, 2005. URL `http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf`.
[20] Patrick Pantel and Marco Pennacchiotti. Esspresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In Calzolari et al. [4], pages 113–120. URL `http://www.aclweb.org/anthology/P/P06/P06-1015`.
[21] Adam Pawłowski, Maciej Piasecki, and Bartosz Broda. Automatic extraction of word-profiles from text corpora. on the example of polish collective symbols. Prepared for the planned post-conference Proceedings of Trewir'07, April 2008.
[22] Maciej Piasecki. Handmade and Automatic Rules for Polish Tagger. In Sojka et al. [34].
[23] Maciej Piasecki and Bartosz Broda. Semantic Similarity Measure of Polish Nouns Based on Linguistic Features. In Witold Abramowicz, editor, *Business Information Systems 10th International Conference, BIS 2007, Poznan, Poland, April 25-27, 2007, Proceedings*, volume 4439 of *Lecture Notes in Computer Science*. Springer, 2007.
[24] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. Extended Similarity Test for the Evaluation of Semantic Similarity Functions. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language and Technology Conference, October 5–7, 2007, Poznań, Poland*, pages 104–108, Poznań, 2007. Wydawnictwo Poznańskie Sp. z o.o.
[25] Maciej Piasecki, Michał Marcińczuk, Stanisław Szpakowicz, and Bartosz Broda. Classification-based Filtering of Semantic Relatedness in Hypernymy Extraction. In *Proceedings of the GoTAL 2008 Conference*, LNAI. Springer, 2008.
[26] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2009. To appear.
[27] Tadeusz Piotrowski and Zygmunt Saloni. *Kieszonkowy słownik angielsko-polski i polsko-angielski*. Wyd. Wilga, Warszawa, 1999.
[28] Adam Przepiórkowski. *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS, 2004.
[29] James Pustejovsky. Generative lexicon. *Computational Linguistics*, 17(4):409–441, 1991.
[30] PWN. Słownik języka polskiego. URL `http://sjp.pwn.pl/`. Published on the web page, 2007.
[31] Rzeczpospolita. Korpus Rzeczpospolitej. [on-line] `www.cs.put.poznan.pl/dweiss/rzeczpospolita`, 2008. Corpus of text from the online edtion of Rzeczpospolita.
[32] Magnus Sahlgren. Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. In *Proceedings of the Semantic Knowledge Acquisition and Categorisation Workshop, ESSLLI 2001*, Helsinki, Finland, 2001.
[33] Rion Snow, Dan Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In Calzolari et al. [4]. URL `http://www.stanford.edu/~jurafsky/COLACL101.pdf`.
[34] Petr Sojka, Ivan Kopecek, and Karel Pala, editors. *Proceedings of the Text, Speech and Dialog 2006 Conference*, LNAI, 2006. Springer.
[35] Anna Wierzbicka. *Semantyka. Jednostki elementarne i uniwersalne*. UMCS, 2006.
[36] Torsten Zesch and Iryna Gurevych. Automatically Creating Datasets for Measures of Semantic Relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24, Sydney, Australia, 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-1104`.