Extracting Lexical Reference Rules from Wikipedia

Eyal Shnarch
Computer Science Department
Bar-Ilan University
Ramat-Gan 52900, Israel
shey@cs.biu.ac.il

Libby Barak
Dept. of Computer Science
University of Toronto
Toronto, Canada M5S 1A4
libbyb@cs.toronto.edu

Ido Dagan
Computer Science Department
Bar-Ilan University
Ramat-Gan 52900, Israel
dagan@cs.biu.ac.il

Abstract

This paper describes the extraction from Wikipedia of *lexical reference* rules, identifying references to term meanings triggered by other terms. We present extraction methods geared to cover the broad range of the lexical reference relation and analyze them extensively. Most extraction methods yield high precision levels, and our rule-base is shown to perform better than other automatically constructed baselines in a couple of lexical expansion and matching tasks. Our rule-base yields comparable performance to WordNet while providing largely complementary information.

1 Introduction

A most common need in applied semantic inference is to infer the meaning of a target term from other terms in a text. For example, a Question Answering system may infer the answer to a question regarding *luxury cars* from a text mentioning *Bentley*, which provides a concrete reference to the sought meaning.

Aiming to capture such lexical inferences we followed (Glickman et al., 2006), which coined the term *lexical reference* (LR) to denote references in text to the specific meaning of a target term. They further analyzed the dataset of the First Recognizing Textual Entailment Challenge (Dagan et al., 2006), which includes examples drawn from seven different application scenarios. It was found that an entailing text indeed includes a concrete reference to practically every term in the entailed (inferred) sentence.

The lexical reference relation between two terms may be viewed as a lexical inference rule, denoted *LHS* $\Rightarrow$ *RHS*. Such rule indicates that the left-hand-side term would generate a reference, in some texts, to a possible meaning of the right hand side term, as the *Bentley* $\Rightarrow$ *luxury car* example.

In the above example the LHS is a hyponym of the RHS. Indeed, the commonly used hyponymy, synonymy and some cases of the meronymy relations are special cases of lexical reference. However, lexical reference is a broader relation. For instance, the LR rule *physician* $\Rightarrow$ *medicine* may be useful to infer the topic *medicine* in a text categorization setting, while an information extraction system may utilize the rule *Margaret Thatcher* $\Rightarrow$ *United Kingdom* to infer a UK announcement from the text "*Margaret Thatcher announced*".

To perform such inferences, systems need large scale knowledge bases of LR rules. A prominent available resource is WordNet (Fellbaum, 1998), from which classical relations such as synonyms, hyponyms and some cases of meronyms may be used as LR rules. An extension to WordNet was presented by (Snow et al., 2006). Yet, available resources do not cover the full scope of lexical reference.

This paper presents the extraction of a large-scale rule base from Wikipedia designed to cover a wide scope of the lexical reference relation. As a starting point we examine the potential of definition sentences as a source for LR rules (Ide and Jean, 1993; Chodorow et al., 1985; Moldovan and Rus, 2001). When writing a concept definition, one aims to formulate a concise text that includes the most characteristic aspects of the defined concept. Therefore, a definition is a promising source for LR relations between the defined concept and the definition terms.

In addition, we extract LR rules from Wikipedia redirect and hyperlink relations. As a guideline, we focused on developing simple extraction methods that may be applicable for other Web knowledge resources, rather than focusing on Wikipedia-specific attributes. Overall, our rule base contains about 8 million candidate lexical ref-

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 450–458,
Suntec, Singapore, 2-7 August 2009.

erence rules. [1]

Extensive analysis estimated that 66% of our rules are correct, while different portions of the rule base provide varying recall-precision tradeoffs. Following further error analysis we introduce rule filtering which improves inference performance. The rule base utility was evaluated within two lexical expansion applications, yielding better results than other automatically constructed baselines and comparable results to WordNet. A combination with WordNet achieved the best performance, indicating the significant marginal contribution of our rule base.

2 Background

Many works on machine readable dictionaries utilized definitions to identify semantic relations between words (Ide and Jean, 1993). Chodorow et al. (1985) observed that the head of the defining phrase is a genus term that describes the defined concept and suggested simple heuristics to find it. Other methods use a specialized parser or a set of regular expressions tuned to a particular dictionary (Wilks et al., 1996).

Some works utilized Wikipedia to build an ontology. Ponzetto and Strube (2007) identified the subsumption (IS-A) relation from Wikipedia's category tags, while in Yago (Suchanek et al., 2007) these tags, redirect links and WordNet were used to identify instances of 14 predefined specific semantic relations. These methods depend on Wikipedia's category system. The lexical reference relation we address subsumes most relations found in these works, while our extractions are not limited to a fixed set of predefined relations.

Several works examined Wikipedia texts, rather than just its structured features. Kazama and Torisawa (2007) explores the first sentence of an article and identifies the first noun phrase following the verb *be* as a label for the article title. We reproduce this part of their work as one of our baselines. Toral and Muñoz (2007) uses all nouns in the first sentence. Gabrilovich and Markovitch (2007) utilized Wikipedia-based concepts as the basis for a high-dimensional meaning representation space.

Hearst (1992) utilized a list of patterns indicative for the hyponym relation in general texts. Snow et al. (2006) use syntactic path patterns as features for supervised hyponymy and synonymy classifiers, whose training examples are derived automatically from WordNet. They use these classifiers to suggest extensions to the WordNet hierarchy, the largest one consisting of 400K new links. Their automatically created resource is regarded in our paper as a primary baseline for comparison.

Many works addressed the more general notion of *lexical associations*, or association rules (e.g. (Ruge, 1992; Rapp, 2002)). For example, *The Beatles*, *Abbey Road* and *Sgt. Pepper* would all be considered lexically associated. However this is a rather loose notion, which only indicates that terms are semantically "related" and are likely to co-occur with each other. On the other hand, lexical reference is a special case of lexical association, which specifies concretely that a reference to the meaning of one term may be inferred from the other. For example, *Abbey Road* provides a concrete reference to *The Beatles*, enabling to infer a sentence like "*I listened to The Beatles*" from "*I listened to Abbey Road*", while it does not refer specifically to *Sgt. Pepper*.

3 Extracting Rules from Wikipedia

Our goal is to utilize the broad knowledge of Wikipedia to extract a knowledge base of lexical reference rules. Each Wikipedia article provides a definition for the concept denoted by the *title* of the article. As the most concise definition we take the first sentence of each article, following (Kazama and Torisawa, 2007). Our preliminary evaluations showed that taking the entire first paragraph as the definition rarely introduces new valid rules while harming extraction precision significantly.

Since a concept definition usually employs more general terms than the defined concept (Ide and Jean, 1993), the concept title is more likely to refer to terms in its definition rather than vice versa. Therefore the title is taken as the LHS of the constructed rule while the extracted definition term is taken as its RHS. As Wikipedia's titles are mostly noun phrases, the terms we extract as RHSs are the nouns and noun phrases in the definition. The remainder of this section describes our methods for extracting rules from the definition sentence and from additional Wikipedia information.

Be-Comp Following the general idea in (Kazama and Torisawa, 2007), we identify the *IS-A* pattern in the definition sentence by extracting nominal complements of the verb 'be', taking

[1] For download see *Textual Entailment Resource Pool* at the ACL-wiki (http://aclweb.org/aclwiki)

No.	Extraction	Rule
James Eugene "Jim" Carrey is a Canadian-American actor and comedian		
1	*Be-Comp*	*Jim Carrey* $\Rightarrow$ *Canadian-American actor*
2	*Be-Comp*	*Jim Carrey* $\Rightarrow$ *actor*
3	*Be-Comp*	*Jim Carrey* $\Rightarrow$ *comedian*
Abbey Road is an album released by The Beatles		
4	*All-N*	*Abbey Road* $\Rightarrow$ *The Beatles*
5	*Parenthesis*	*Graph* $\Rightarrow$ *mathematics*
6	*Parenthesis*	*Graph* $\Rightarrow$ *data structure*
7	*Redirect*	*CPU* $\Leftrightarrow$ *Central processing unit*
8	*Redirect*	*Receptors IgG* $\Leftrightarrow$ *Antibody*
9	*Redirect*	*Hypertension* $\Leftrightarrow$ *Elevated blood-pressure*
10	*Link*	*pet* $\Rightarrow$ *Domesticated Animal*
11	*Link*	*Gestaltist* $\Rightarrow$ *Gestalt psychology*

Table 1: Examples of rule extraction methods

them as the RHS of a rule whose LHS is the article title. While Kazama and Torisawa used a chunker, we parsed the definition sentence using Minipar (Lin, 1998b). Our initial experiments showed that parse-based extraction is more accurate than chunk-based extraction. It also enables us extracting additional rules by splitting conjoined noun phrases and by taking both the head noun and the complete base noun phrase as the RHS for separate rules (examples 1–3 in Table 1).

All-N The *Be-Comp* extraction method yields mostly hypernym relations, which do not exploit the full range of lexical references within the concept definition. Therefore, we further create rules for all head nouns and base noun phrases within the definition (example 4). An unsupervised reliability score for rules extracted by this method is investigated in Section 4.3.

Title Parenthesis A common convention in Wikipedia to disambiguate ambiguous titles is adding a descriptive term in parenthesis at the end of the title, as in *The Siren (Musical)*, *The Siren (sculpture)* and *Siren (amphibian)*. From such titles we extract rules in which the descriptive term inside the parenthesis is the RHS and the rest of the title is the LHS (examples 5–6).

Redirect As any dictionary and encyclopedia, Wikipedia contains *Redirect* links that direct different search queries to the same article, which has a canonical title. For instance, there are 86 different queries that redirect the user to *United States* (e.g. *U.S.A.*, *America*, *Yankee land*). Redirect links are hand coded, specifying that both terms refer to the same concept. We therefore generate a bidirectional entailment rule for each redirect link (examples 7–9).

Link Wikipedia texts contain hyper links to articles. For each link we generate a rule whose LHS is the linking text and RHS is the title of the linked article (examples 10–11). In this case we generate a directional rule since links do not necessarily connect semantically equivalent entities.

We note that the last three extraction methods should not be considered as Wikipedia specific, since many Web-like knowledge bases contain redirects, hyper-links and disambiguation means. Wikipedia has additional structural features such as category tags, structured summary tablets for specific semantic classes, and articles containing lists which were exploited in prior work as reviewed in Section 2.

As shown next, the different extraction methods yield different precision levels. This may allow an application to utilize only a portion of the rule base whose precision is above a desired level, and thus choose between several possible recall-precision tradeoffs.

4 Extraction Methods Analysis

We applied our rule extraction methods over a version of Wikipedia available in a database constructed by (Zesch et al., 2007)[2]. The extraction yielded about 8 million rules altogether, with over 2.4 million distinct RHSs and 2.8 million distinct LHSs. As expected, the extracted rules involve mostly named entities and specific concepts, typically covered in encyclopedias.

4.1 Judging Rule Correctness

Following the spirit of the fine-grained human evaluation in (Snow et al., 2006), we randomly sampled 800 rules from our rule-base and presented them to an annotator who judged them for correctness, according to the lexical reference notion specified above. In cases which were too difficult to judge the annotator was allowed to abstain, which happened for 20 rules. 66% of the remaining rules were annotated as correct. 200 rules from the sample were judged by another annotator for agreement measurement. The resulting Kappa score was 0.7 (substantial agreement (Landis and

[2]English version from February 2007, containing 1.6 million articles. www.ukp.tu-darmstadt.de/software/JWPL

Extraction Method	Per Method		Accumulated	
	P	Est. #Rules	P	%obtained
Redirect	0.87	1,851,384	0.87	31
Be-Comp	0.78	1,618,913	0.82	60
Parenthesis	0.71	94,155	0.82	60
Link	0.7	485,528	0.80	68
All-N	0.49	1,580,574	0.66	100

Table 2: Manual analysis: precision and estimated number of correct rules per extraction method, and precision and % of correct rules obtained of rule-sets accumulated by method.

Koch, 1997)), either when considering all the abstained rules as correct or as incorrect.

The middle columns of Table 2 present, for each extraction method, the obtained percentage of correct rules (precision) and their estimated absolute number. This number is estimated by multiplying the number of annotated correct rules for the extraction method by the sampling proportion. In total, we estimate that our resource contains 5.6 million correct rules. For comparison, Snow's published extension to WordNet[3], which covers similar types of terms but is restricted to synonyms and hyponyms, includes 400,000 relations.

The right part of Table 2 shows the performance figures for accumulated rule bases, created by adding the extraction methods one at a time in order of their precision. *% obtained* is the percentage of correct rules in each rule base out of the total number of correct rules extracted jointly by all methods (the union set).

We can see that excluding the *All-N* method all extraction methods reach quite high precision levels of 0.7-0.87, with accumulated precision of 0.8[4]. By selecting only a subset of the extraction methods, according to their precision, one can choose different recall-precision tradeoff points that suit application preferences.

The less accurate *All-N* method may be used when high recall is important, accounting for 32% of the correct rules. An examination of the paths in *All-N* reveals, beyond standard hyponymy and synonymy, various semantic relations that satisfy lexical reference, such as *Location*, *Occupation* and *Creation*, as illustrated in Table 3. Typical relations covered by *Redirect* and *Link* rules include synonyms (*NY State Trooper* ⇒ *New York State Police*), morphological derivations (*irritate* ⇒ *irritation*), different spellings or naming (*Pytagoras* ⇒ *Pythagoras*) and acronyms (*AIS* ⇒ *Alarm Indication Signal*).

[3]http://ai.stanford.edu/∼rion/swn/

[4]As a non-comparable reference, Snow's fine-grained evaluation showed a precision of 0.84 on 10K rules and 0.68 on 20K rules; however, they were interested only in the hyponym relation while we evaluate our rules according to the broader LR relation.

4.2 Error Analysis

We sampled 100 rules which were annotated as incorrect and examined the causes of errors. Figure 1 shows the distribution of error types.

Wrong NP part - The most common error (35% of the errors) is taking an inappropriate part of a noun phrase (NP) as the rule right hand side (RHS). As described in Section 3, we create two rules from each extracted NP, by taking both the head noun and the complete base NP as RHSs. While both rules are usually correct, there are cases in which the left hand side (LHS) refers to the NP as a whole but not to part of it. For example, *Margaret Thatcher* refers to *United Kingdom* but not to *Kingdom*. In Section 5 we suggest a filtering method which addresses some of these errors. Future research may exploit methods for detecting multi-words expressions.

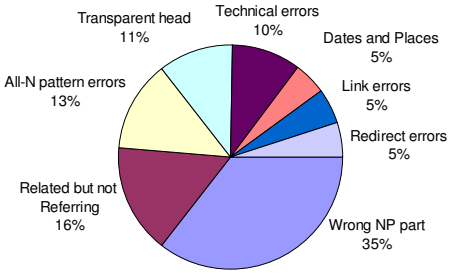

Figure 1: Error analysis: type of incorrect rules

Related but not Referring - Although all terms in a definition are highly related to the defined concept, not all are referred by it. For example the origin of a person (**The Beatles* ⇒ *Liverpool*[5]) or family ties such as 'daughter of' or 'sire of'.

All-N errors - Some of the articles start with a long sentence which may include information that is not directly referred by the title of the article. For instance, consider **Interstate 80* ⇒ *California* from "*Interstate 80 runs from California to New Jersey*". In Section 4.3 we further analyze this type of error and point at a possible direction for addressing it.

Transparent head - This is the phenomenon in which the syntactic head of a noun phrase does

[5]The asterisk denotes an incorrect rule

Relation	Rule	Path Pattern
Location	*Lovek* ⇒ *Cambodia*	*Lovek* city in *Cambodia*
Occupation	*Thomas H. Cormen* ⇒ *computer science*	*Thomas H. Cormen* professor of *computer science*
Creation	*Genocidal Healer* ⇒ *James White*	*Genocidal Healer* novel by *James White*
Origin	*Willem van Aelst* ⇒ *Dutch*	*Willem van Aelst Dutch* artist
Alias	*Dean Moriarty* ⇒ *Benjamin Linus*	*Dean Moriarty* is an alias of *Benjamin Linus* on Lost.
Spelling	*Egushawa* ⇒ *Agushaway*	*Egushawa*, also spelled *Agushaway*...

Table 3: *All-N* rules exemplifying various types of LR relations

not bear its primary meaning, while it has a modifier which serves as the semantic head (Fillmore et al., 2002; Grishman et al., 1986). Since parsers identify the syntactic head, we extract an incorrect rule in such cases. For instance, deriving **Prince William ⇒ member* instead of *Prince William ⇒ British Royal Family* from "*Prince William is a member of the British Royal Family*". Even though we implemented the common solution of using a list of typical transparent heads, this solution is partial since there is no closed set of such phrases.

Technical errors - Technical extraction errors were mainly due to erroneous identification of the title in the definition sentence or mishandling non-English texts.

Dates and Places - Dates and places where a certain person was born at, lived in or worked at often appear in definitions but do not comply to the lexical reference notion (**Galileo Galilei ⇒ 15 February 1564*).

Link errors - These are usually the result of wrong assignment of the reference direction. Such errors mostly occur when a general term, e.g. *revolution*, links to a more specific albeit typical concept, e.g. *French Revolution*.

Redirect errors - These may occur in some cases in which the extracted rule is not bidirectional. E.g. **Anti-globalization ⇒ Movement of Movements* is wrong but the opposite entailment direction is correct, as *Movement of Movements* is a popular term in Italy for *Anti-globalization*.

4.3 Scoring All-N Rules

We observed that the likelihood of nouns mentioned in a definition to be referred by the concept title depends greatly on the syntactic path connecting them (which was exploited also in (Snow et al., 2006)). For instance, the path produced by Minipar for example 4 in Table 1 is *title* $\overset{subj}{\longleftarrow}$album$\overset{vrel}{\longrightarrow}$released $\overset{by-subj}{\longrightarrow}by\overset{pcomp-n}{\longrightarrow}$ *noun*.

In order to estimate the likelihood that a syntactic path indicates lexical reference we collected from Wikipedia all paths connecting a title to a noun phrase in the definition sentence. We note that since there is no available resource which covers the full breadth of lexical reference we could not obtain sufficiently broad supervised training data for learning which paths correspond to correct references. This is in contrast to (Snow et al., 2005) which focused only on hyponymy and synonymy relations and could therefore extract positive and negative examples from WordNet.

We therefore propose the following unsupervised reference likelihood score for a syntactic path p within a definition, based on two counts: the number of times p connects an article *title* with a *noun* in its definition, denoted by $C_t(p)$, and the total number of p's occurrences in Wikipedia definitions, $C(p)$. The score of a path is then defined as $\frac{C_t(p)}{C(p)}$. The rational for this score is that $C(p) - C_t(p)$ corresponds to the number of times in which the path connects two nouns within the definition, none of which is the title. These instances are likely to be non-referring, since a concise definition typically does not contain terms that can be inferred from each other. Thus our score may be seen as an approximation for the probability that the two nouns connected by an arbitrary occurrence of the path would satisfy the reference relation. For instance, the path of example 4 obtained a score of 0.98.

We used this score to sort the set of rules extracted by the *All-N* method and split the sorted list into 3 thirds: *top*, *middle* and *bottom*. As shown in Table 4, this obtained reasonably high precision for the top third of these rules, relative to the other two thirds. This precision difference indicates that our unsupervised path score provides useful information about rule reliability.

It is worth noting that in our sample 57% of *All-N* errors, 62% of *Related but not Referring* incorrect rules and all incorrect rules of type *Dates and*

Extraction Method	Per Method		Accumulated	
	P	Est. #Rules	P	%obtained
All-N_{top}	0.60	684,238	0.76	83
All-N_{middle}	0.46	380,572	0.72	90
All-N_{bottom}	0.41	515,764	0.66	100

Table 4: Splitting *All-N* extraction method into 3 sub-types. These three rows replace the last row of Table 2

Places were extracted by the *All-N_{bottom}* method and thus may be identified as less reliable. However, this split was not observed to improve performance in the application oriented evaluations of Section 6. Further research is thus needed to fully exploit the potential of the syntactic path as an indicator for rule correctness.

5 Filtering Rules

Following our error analysis, future research is needed for addressing each specific type of error. However, during the analysis we observed that all types of erroneous rules tend to relate terms that are rather unlikely to co-occur together. We therefore suggest, as an optional filter, to recognize such rules by their co-occurrence statistics using the common Dice coefficient:

$$\frac{2 \cdot C(LHS, RHS)}{C(LHS) + C(RHS)}$$

where $C(x)$ is the number of articles in Wikipedia in which all words of x appear.

In order to partially overcome the *Wrong NP part* error, identified in Section 4.2 to be the most common error, we adjust the Dice equation for rules whose RHS is also part of a larger noun phrase (NP):

$$\frac{2 \cdot (C(LHS, RHS) - C(LHS, NP_{RHS}))}{C(LHS) + C(RHS)}$$

where NP_{RHS} is the complete NP whose part is the *RHS*. This adjustment counts only co-occurrences in which the LHS appears with the RHS alone and not with the larger NP. This substantially reduces the Dice score for those cases in which the LHS co-occurs mainly with the full NP.

Given the Dice score rules whose score does not exceed a threshold may be filtered. For example, the incorrect rule **aerial tramway* $\Rightarrow$ *car* was filtered, where the correct RHS for this LHS is the complete NP *cable car*. Another filtered rule is *magic* $\Rightarrow$ *cryptography* which is correct only for a very idiosyncratic meaning.[6]

We also examined another filtering score, the cosine similarity between the vectors representing the two rule sides in LSA (Latent Semantic Analysis) space (Deerwester et al., 1990). However, as the results with this filter resemble those for Dice we present results only for the simpler Dice filter.

6 Application Oriented Evaluations

Our primary application oriented evaluation is within an unsupervised lexical expansion scenario applied to a text categorization data set (Section 6.1). Additionally, we evaluate the utility of our rule base as a lexical resource for recognizing textual entailment (Section 6.2).

6.1 Unsupervised Text Categorization

Our categorization setting resembles typical query expansion in information retrieval (IR), where the category name is considered as the query. The advantage of using a text categorization test set is that it includes exhaustive annotation for *all* documents. Typical IR datasets, on the other hand, are partially annotated through a pooling procedure. Thus, some of our valid lexical expansions might retrieve non-annotated documents that were missed by the previously pooled systems.

6.1.1 Experimental Setting

Our categorization experiment follows a typical keywords-based text categorization scheme (McCallum and Nigam, 1999; Liu et al., 2004). Taking a lexical reference perspective, we assume that the characteristic expansion terms for a category should refer to the term (or terms) denoting the category name. Accordingly, we construct the category's feature vector by taking first the category name itself, and then expanding it with all left-hand sides of lexical reference rules whose right-hand side is the category name. For example, the category "Cars" is expanded by rules such as *Ferrari F50* $\Rightarrow$ *car*. During classification cosine similarity is measured between the feature vector of the classified document and the expanded vectors of all categories. The document is assigned to the category which yields the highest similarity score, following a single-class classification approach (Liu et al., 2004).

[6] Magic was the United States codename for intelligence derived from cryptanalysis during World War II.

Rule Base	R	P	F_1
Baselines:			
No Expansion	0.19	0.54	0.28
WikiBL	0.19	0.53	0.28
Snow^{400K}	0.19	0.54	0.28
Lin	0.25	0.39	0.30
WordNet	0.30	0.47	0.37
Extraction Methods from Wikipedia:			
Redirect + Be-Comp	0.22	**0.55**	0.31
All rules	0.31	0.38	0.34
All rules + Dice filter	0.31	0.49	**0.38**
Union:			
WordNet + Wiki$_{All\_rules+Dice}$	**0.35**	0.47	**0.40**

Table 5: Results of different rule bases for 20 newsgroups category name expansion

It should be noted that keyword-based text categorization systems employ various additional steps, such as bootstrapping, which generalize to multi-class settings and further improve performance. Our basic implementation suffices to evaluate comparatively the direct impact of different expansion resources on the initial classification.

For evaluation we used the test set of the "bydate" version of the 20-News Groups collection,[7] which contains 18,846 documents partitioned (nearly) evenly over the 20 categories[8].

6.1.2 Baselines Results

We compare the quality of our rule base expansions to 5 baselines (Table 5). The first avoids any expansion, classifying documents based on cosine similarity with category names only. As expected, it yields relatively high precision but low recall, indicating the need for lexical expansion.

The second baseline is our implementation of the relevant part of the Wikipedia extraction in (Kazama and Torisawa, 2007), taking the first noun after a *be* verb in the definition sentence, denoted as *WikiBL*. This baseline does not improve performance at all over no expansion.

The next two baselines employ state-of-the-art lexical resources. One uses Snow's extension to WordNet which was mentioned earlier. This resource did not yield a noticeable improvement, either over the *No Expansion* baseline or over *WordNet* when joined with its expansions. The second uses *Lin* dependency similarity, a syntactic-dependency based distributional word similarity resource described in (Lin, 1998a)[9]. We used various thresholds on the length of the expansion list derived from this resource. The best result, reported here, provides only a minor F_1 improvement over *No Expansion*, with modest recall increase and significant precision drop, as can be expected from such distributional method.

The last baseline uses *WordNet* for expansion. First we expand all the senses of each category name by their derivations and synonyms. Each obtained term is then expanded by its hyponyms, or by its meronyms if it has no hyponyms. Finally, the results are further expanded by their derivations and synonyms.[10] *WordNet* expansions improve substantially both Recall and F_1 relative to *No Expansion*, while decreasing precision.

6.1.3 Wikipedia Results

We then used for expansion different subsets of our rule base, producing alternative recall-precision tradeoffs. Table 5 presents the most interesting results. Using any subset of the rules yields better performance than any of the other automatically constructed baselines (*Lin*, *Snow* and *WikiBL*). Utilizing the most precise extraction methods of *Redirect* and *Be-Comp* yields the highest precision, comparable to *No Expansion*, but just a small recall increase. Using the entire rule base yields the highest recall, while filtering rules by the Dice coefficient (with 0.1 threshold) substantially increases precision without harming recall. With this configuration our automatically-constructed resource achieves comparable performance to the manually built *WordNet*.

Finally, since a dictionary and an encyclopedia are complementary in nature, we applied the union of *WordNet* and the filtered *Wikipedia* expansions. This configuration yields the best results: it maintains *WordNet*'s precision and adds nearly 50% to the recall increase of *WordNet* over *No Expansion*, indicating the substantial marginal contribution of *Wikipedia*. Furthermore, with the fast growth of Wikipedia the recall of our resource is expected to increase while maintaining its precision.

[7] www.ai.mit.edu/people/jrennie/20Newsgroups.

[8] The keywords used as category names are: atheism; graphic; microsoft windows; ibm,pc,hardware; mac,hardware; x11,x-windows; sale; car; motorcycle; baseball; hockey; cryptography; electronics; medicine; outer space; christian(noun & adj); gun; mideast,middle east; politics; religion

[9] Downloaded from www.cs.ualberta.ca/lindek/demos.htm

[10] We also tried expanding by the entire hyponym hierarchy and considering only the first sense of each synset, but the method described above achieved the best performance.

Category Name	Expanding Terms
Politics	opposition, coalition, whip$^{(a)}$
Cryptography	adversary, cryptosystem, key
Mac	PowerBook, Radius$^{(b)}$, Grab$^{(c)}$
Religion	heaven, creation, belief, missionary
Medicine	doctor, physician, treatment, clinical
Computer Graphics	radiosity$^{(d)}$, rendering, siggraph$^{(e)}$

Table 6: Some *Wikipedia* rules not in *WordNet*, which contributed to text categorization. (a) a legislator who enforce leadership desire (b) a hardware firm specializing in Macintosh equipment (c) a Macintosh screen capture software (d) an illumination algorithm (e) a computer graphics conference

Configuration	Accuracy	Accuracy Drop
WordNet + Wikipedia	60.0 %	-
Without WordNet	57.7 %	2.3 %
Without Wikipedia	58.9 %	1.1 %

Table 7: RTE accuracy results for ablation tests.

Table 6 illustrates few examples of useful rules that were found in *Wikipedia* but not in *WordNet*. We conjecture that in other application settings the rules extracted from Wikipedia might show even greater marginal contribution, particularly in specialized domains not covered well by WordNet. Another advantage of a resource based on Wikipedia is that it is available in many more languages than WordNet.

6.2 Recognizing Textual Entailment (RTE)

As a second application-oriented evaluation we measured the contributions of our (filtered) Wikipedia resource and WordNet to RTE inference (Giampiccolo et al., 2007). To that end, we incorporated both resources within a typical basic RTE system architecture (Bar-Haim et al., 2008). This system determines whether a text entails another sentence based on various matching criteria that detect syntactic, logical and lexical correspondences (or mismatches). Most relevant for our evaluation, lexical matches are detected when a Wikipedia rule's LHS appears in the text and its RHS in the hypothesis, or similarly when pairs of WordNet synonyms, hyponyms-hypernyms and derivations appear across the text and hypothesis. The system's weights were trained on the development set of RTE-3 and tested on RTE-4 (which included this year only a test set).

To measure the marginal contribution of the two resources we performed ablation tests, comparing the accuracy of the full system to that achieved when removing either resource. Table 7 presents the results, which are similar in nature to those obtained for text categorization. *Wikipedia* obtained a marginal contribution of 1.1%, about half of the analogous contribution of WordNet's manually-constructed information. We note that for current RTE technology it is very typical to gain just a few percents in accuracy thanks to external knowledge resources, while individual resources usually contribute around 0.5–2% (Iftene and Balahur-Dobrescu, 2007; Dinu and Wang, 2009). Some *Wikipedia* rules not in WordNet which contributed to RTE inference are *Jurassic Park* $\Rightarrow$ *Michael Crichton*, *GCC* $\Rightarrow$ *Gulf Cooperation Council*.

7 Conclusions and Future Work

We presented construction of a large-scale resource of lexical reference rules, as useful in applied lexical inference. Extensive rule-level analysis showed that different recall-precision tradeoffs can be obtained by utilizing different extraction methods. It also identified major reasons for errors, pointing at potential future improvements. We further suggested a filtering method which significantly improved performance.

Even though the resource was constructed by quite simple extraction methods, it was proven to be beneficial within two different application setting. While being an automatically built resource, extracted from a knowledge-base created for human consumption, it showed comparable performance to WordNet, which was manually created for computational purposes. Most importantly, it also provides complementary knowledge to WordNet, with unique lexical reference rules.

Future research is needed to improve resource's precision, especially for the *All-N* method. As a first step, we investigated a novel unsupervised score for rules extracted from definition sentences. We also intend to consider the rule base as a directed graph and exploit the graph structure for further rule extraction and validation.

Acknowledgments

The authors would like to thank Idan Szpektor for valuable advices. This work was partially supported by the NEGEV project (www.negev-initiative.org), the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and by the Israel Science Foundation grant 1112/08.

References

Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of TAC*.

Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of ACL*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Lecture Notes in Computer Science*, volume 3944, pages 177–190.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Georgiana Dinu and Rui Wang. 2009. Inference rules for recognizing textual entailment. In *Proceedings of the IWCS*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. Seeing arguments through transparent structures. In *Proceedings of LREC*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of ACL-WTEP Workshop*.

Oren Glickman, Eyal Shnarch, and Ido Dagan. 2006. Lexical reference: a semantic matching subtask. In *Proceedings of EMNLP*.

Ralph Grishman, Lynette Hirschman, and Ngo Thanh Nhan. 1986. Discovery procedures for sublanguage selectional patterns: Initial experiments. *Computational Linguistics*, 12(3):205–215.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*.

Nancy Ide and Véronis Jean. 1993. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In *Proceedings of KB & KS Workshop*.

Adrian Iftene and Alexandra Balahur-Dobrescu. 2007. Hypothesis transformation and semantic variability rules used in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.

J. Richard Landis and Gary G. Koch. 1997. The measurements of observer agreement for categorical data. In *Biometrics*, pages 33:159–174.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

Dekang Lin. 1998b. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC*.

Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. 2004. Text classification by labeling words. In *Proceedings of AAAI*.

Andrew McCallum and Kamal Nigam. 1999. Text classification by bootstrapping with keywords, EM and shrinkage. In *Proceedings of ACL Workshop for unsupervised Learning in NLP*.

Dan Moldovan and Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of ACL*.

Simone P. Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from wikipedia. In *Proceedings of AAAI*.

Reinhard Rapp. 2002. The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Proceedings of COLING*.

Gerda Ruge. 1992. Experiment on linguistically-based term associations. *Information Processing & Management*, 28(3):317–332.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING-ACL*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge - unifying wordnet and wikipedia. In *Proceedings of WWW*.

Antonio Toral and Rafael Muñoz. 2007. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of NAACL/HLT*.

Yorick A. Wilks, Brian M. Slator, and Louise M. Guthrie. 1996. *Electric words: dictionaries, computers, and meanings*. MIT Press, Cambridge, MA, USA.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing wikipedia as a lexical semantic resource. In *Data Structures for Linguistic Resources and Applications*, pages 197–205.