Do wordnets also improve human performance on NLP tasks?

Kristiina Muhonen and **Krister Lindén**
Department of Modern Languages
University of Helsinki
kristiina.muhonen@helsinki.fi krister.linden@helsinki.fi

Abstract

FinnWordNet is a wordnet for Finnish that complies with the format of the Princeton WordNet (PWN) (Fellbaum, 1998). It was built by translating the Princeton WordNet 3.0 synsets into Finnish by human translators. It is open source and contains 117000 synsets. The Finnish translations were inserted into the PWN structure resulting in a bilingual lexical database.

In natural language processing (NLP), wordnets have been used for infusing computers with semantic knowledge assuming that humans already have a sufficient amount of this knowledge.

In this paper we present a case study of using wordnets as an electronic dictionary. We tested whether native Finnish speakers benefit from using a wordnet while completing English sentence completion tasks. We found that using either an English wordnet or a bilingual English-Finnish wordnet significantly improves performance in the task. This should be taken into account when setting standards and comparing human and computer performance on these tasks.

1 Introduction

Wordnets are lexical databases that group words of a language into synonym sets called synsets, provide general definitions of the synsets and encode the semantic relations between the synsets. Typically they are monolingual, but efforts have been made to produce multilingual wordnets as well, see e.g. Vossen (1998).

1.1 Building a New Wordnet

A wordnet for a new language can be constructed in several ways. First, it can be built from scratch. This requires extracting the synsets automatically from corpora or defining them manually. In order to ensure that the most common words of the language are actually present in the automatically collected synsets, a common strategy is to use a list with the central vocabulary of the language. Not only do the actual synsets need to be automatically extracted, also the semantic relations between the synsets must be encoded from the very beginning.

Second, the new wordnet can be translated from an existing wordnet. Translating a wordnet ignores the idea of every language being so different with such varying synonym groups and hierarchies that they have to be constructed separately for every language. However, like Lindén and Carlson (2010) note, most of the words in a language actually describe entities and phenomena present in most languages, although there are language specific differences in which nuances of a concept get a specific word capturing the distinctions in meaning.

The third way to construct a wordnet is a combination of automatic extraction and translation. First, the core of the new wordnet is built by translating 5000 central concepts from the PWN. This core can be extended with e.g. a thesaurus of the target language. Vossen (2004) describes how basing the wordnet on a common core enables linking wordnets to each other via the *Inter-Lingual-Index*.

1.2 The Finnish Wordnet

FinnWordNet (FiWN) is a direct translation of the synsets in the PWN 3.0. Choosing translation as the way to create the wordnet is motivated by the benefits it brings. Direct translation of an already existing wordnet results in a parallel arrangement of the synsets. This directly provides us with a wordnet that can be used as a bilingual dictionary. Also, most of the semantic relations from the PWN can be directly used in FiWN (Lindén

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 146–152

and Carlson, 2010).

Choosing translation as the means of building FiWN has the downside of including many English-specific terms and concepts in the Finnish wordnet. However, English-specific or rare words in general are all welcome in an electronic version. Some synsets may seem problematic from a cultural perspective, e.g. *independence day* as synonymous with *4th of July*. In such cases, the less general concept can be made a hyponym with corresponding culture-specific terms as sister concepts.

1.3 Using Wordnets

Generally the research around wordnets revolves around NLP applications and less emphasis has been put on wordnets aiding human users. The usability of wordnets as lexical resources for NLP applications has long been established. In particular, wordnets have been found useful in improving the performance of systems for word sense disambiguation, information retrieval and automatic text classification, see e.g. Tanács et al. (2007).

The usability of wordnets for human users is a rather neglected topic. Since creating a wordnet consumes a lot of time and resources, the usability of wordnets should also be considered from a human perspective. The benefits humans get from an intuitively structured lexical database should be considered a prerequisite, not merely a positive side effect of the various wordnet projects implemented for different languages.

The focus of our study is to examine the usability of wordnets from a human perspective. We want to see how human users benefit from using wordnets as a lexical resource and compare the benefits they get from a regular electronic dictionary and first, a monolingual wordnet, and second, a bilingual one. We want to demonstrate that even a monolingual wordnet helps a non-native English speaker complete a sentence completion task at least as much as a regular dictionary does.

2 Method

The purpose of our study is to examine how wordnets aid human users. The experiment is conducted by asking Finnish native speakers to carry out sentence completion tasks in English using different lexical resources for assistance.

The test consists of SAT Reasoning Test style multiple choice sentence completion tasks. We decided on using sentence completion rather than translation because it is more straightforward to assess the correctness of the answers. Had we chosen a translation task, we would have first had to decide what the ultimately best translation of a given English passage is, which is a complex issue reaching far beyond the scope of this paper.

We sought out the sample questions from a set of training questions for the SAT test[1]. We estimated that SAT-level English is sufficiently difficult for Finnish university students, so that the testees would not be able to get full scores on the test without using any help.

Sentence completions measure the testees vocabulary and understanding of sentence structure and require the testee to select one or two words that best complete the sentence. The questions are multiple choice and there are five options to choose from. In Figure 1 we display an example question from the test.

1. ___ by nature, Jones spoke very little even to his own family members.
A. garrulous
B. equivocal
C. taciturn
D. arrogant
E. gregarious

Figure 1: A sample question

There are 40 questions randomly grouped into sets of ten. Each set is completed with the help of a different aid.

2.1 Lexical Resources

The purpose of the experiment is to see which lexical aid helps the testee the most. In order to see this, we ask the testees to use three different tools: an electronic English dictionary, PWN and FiWN. One set of questions is answered without using any help so that we can establish the English vocabulary skills of the answerer.

2.1.1 Merriam-Webster

We chose the Merriam-Webster[2] English dictionary as the electronic dictionary since it is widely used and freely available. First off we thought of using an English-Finnish dictionary, but since the task is not about translation, the English dictionary better suits our needs. Using an English dictionary we can see what help the testees receive

[1] http://www.majortests.com/sat/sentence-completion.php
[2] http://www.merriam-webster.com

from a regular dictionary without Finnish translations. We assume this to be the most typical kind of lexical aid used. An abridged dictionary entry for "equivocal"[3] can be seen in Figure 2 with the boldface words being links to other entries.

equiv·o·cal
1.
a: subject to two or more **interpretations** and usually used to mislead or confuse <an equivocal statement>
b: uncertain as an indication or sign <equivocal evidence>
2.
a: of uncertain nature or classification <equivocal shapes>
b: of uncertain **disposition** toward a person or thing : **undecided** <an equivocal attitude>
c: of doubtful advantage, genuineness, or moral **rectitude** <equivocal behavior>
Examples of EQUIVOCAL
He responded to reporters' questions with *equivocal* answers.
The experiment produced *equivocal* results.

Figure 2: A truncated Merriam-Webster dictionary entry

Merriam-Webster also includes information about the etymology of the word and a list of synonyms and antonyms. For conciseness sake we do not repeat the information here.

2.1.2 PWN

The second resource we use is the PWN. [4] We give a truncated PWN search result for the word *equivocal*[5] in Figure 3.

Overview of adj equivocal
The adj equivocal has 3 senses (first 1 from tagged texts)
1. (1) equivocal, **ambiguous** – (open to two or more interpretations; or of uncertain nature or significance; or (often) intended to mislead; "an equivocal statement"; [...])
2. equivocal – (open to question; "aliens of equivocal loyalty"; [...])
3. equivocal – (uncertain as a sign or indication; "the evidence from bacteriologic analysis was equivocal")

Figure 3: A truncated PWN entry: Overview

If the user only uses the "Overview" mode of the PWN, the use of the wordnet resembles that of a regular dictionary. Only when the user also views the "Similarity" information of the word, does the structure of the wordnet benefit the user. This can be seen in Figure 4. The boldfaced words are again links to other entries.

Similarity of adj *equivocal*
Sense 1 *equivocal* (vs. **unequivocal**), **ambiguous**
=> **double**, **forked**
=> **evasive**
=> **indeterminate**
Also See-> **ambiguous**#2
Sense 2 *equivocal*
=> **questionable** (vs. **unquestionable**)
Sense 3 *equivocal*
=> **inconclusive** (vs. **conclusive**)

Figure 4: A truncated PWN entry: Similarity

It is also possible to view e.g. antonyms, pertainyms, derived forms and the polysemy count of the word, but for the task at hand the information presented in Figures 4 and 3 suffices.

We assume that using the English wordnet yields at least slightly better results in the sentence completion task than using an electronic dictionary. The assumption based on the intuitive grouping of the synsets.

2.1.3 FiWN

The third tool to be used in the test is the PWN with the Finnish translations visible, FiWN[6]. The search results are identical to the PWN, only the Finnish translations are added. The glosses and examples are still only in English. The overview of the translated adjective *equivocal* is shown in Figure 5.

Overview of adj equivocal
The adj equivocal has 3 senses (first 1 from tagged texts)
1. (1) equivocal [**kaksiselitteinen**], **ambiguous** [**monikäsitteinen**, **epäselvä**, **monimerkityksinen**] – (open to two or more interpretations; or of uncertain nature or significance; or (often) intended to mislead; "an equivocal statement"; [...])
2. equivocal [**epävarma**, **kyseenalainen**] – (open to question; "aliens of equivocal loyalty"; [...])
3. equivocal [**epävarma**, **kyseenalainen**] – (uncertain as a sign or indication; "the evidence from bacteriologic analysis was equivocal")

Figure 5: A truncated FiWN entry with translations: Overview

Correspondingly, Figure 6 shows the synsets with the Finnish equivalents.

We want to see whether the results get significantly better when the testee gets to use a bilingual wordnet. At first guess it can be assumed that the translations speed up the test taking and improve

[3] http://www.merriam-webster.com/dictionary/equivocal
[4] http://wordnet.princeton.edu/
[5] http://www.ling.helsinki.fi/cgi-bin/fiwn/search?wn=en&w=equivocal&t=all&sm=Search
[6] http://www.ling.helsinki.fi/cgi-bin/fiwn/search?

Similarity of adj *equivocal*
Sense 1 *equivocal* [**kaksiselitteinen**] (vs. **unequivocal** [**vastaansanomaton**, **selkeä**]), **ambiguous** [**monikäsitteinen**, **epäselvä**, **monimerkityksinen**]
=> **double** [**kaksimielinen**], **forked** [**kaksimielinen**]
=> **evasive** [**välttelevä**, **kartteleva**]
=> **indeterminate** [**epämääräinen**]
Also See-> **ambiguous**#2 [**monikäsitteinen**, **epäselvä**, **monimerkityksinen**]
Sense 2 *equivocal* [**epävarma**, **kyseenalainen**]
=> **questionable** [**kyseenalainen**] (vs. **unquestionable** [**kiistaton**])
Sense 3 *equivocal* [**epävarma**, **kyseenalainen**]
=> **inconclusive** [**ei ratkaiseva**] (vs. **conclusive** [**ratkaiseva**])

Figure 6: A truncated FiWN entry with translations: Similarity

the results. Since we do not time the test taking, it is only possible to see whether the results get better.

2.2 The Test in Practice

We want to make sure that the randomly chosen questions are equally difficult and that the results are not influenced by one set of questions being easier or harder than the other. To ensure this, we circulate the tool used for each group as shown in Table 1.

QUESTIONS	1-10	11-20	21-30	31-40
TOOL USED	∅	M-W	PWN	FiWN
	FiWN	∅	M-W	PWN
	PWN	FiWN	∅	M-W
	M-W	PWN	FiWN	∅

∅= NO HELP
M-W= MERRIAM-WEBSTER

Table 1: Tool circulation

The test is conducted as an online query. The questions are organized in e-forms which are divided into four parts depending on what type of help the answerer can use. Due to the tool circulation, there are four different e-forms, the order of the tools corresponding to the lines in Table 1.

The test is conducted without supervision or timing. The lack of supervision is due to practical issues; the number of answers is higher when the testees can complete the task whenever it suits them best. This, however, means that the results can be faked. In order to make cheating in the task less tempting, the test is submitted anonymously.

3 Results

We got 34 responses to our query during three weeks with only one reminder. Though the number of testees is fairly small, we can still make general remarks on the usefulness of the three lexical aids as well as on their statistical significance.

Based on the 34 answers we can show that even using an English dictionary significantly improves the performance of the testee. This is a rather predictable outcome. The more interesting question is, whether using PWN as a dictionary improves the results further. And finally, whether a bilingual English-Finnish wordnet brings any further assistance compared to the English one.

Table 2 summarizes the results by showing the average of correct answers per tool. The maximum score is 10.

TOOL	MEAN	MEDIAN	MODE
∅	6.99	7	8
M-W	8.57	9	9
PWN	8.91	9	9
FWN	8.73	9	9

Table 2: Results per tool

Table 2 shows how different tools help users in completing the task. At first look it can already be seen that using any of the chosen tools improves the results, and that the difference between the tools is small.

From Table 3 we can deduce that the difficulty level of the groups is relatively even although the third group seems to have been slightly harder than the rest.

Based on this sample, we cannot draw any conclusions on whether the order of the tools used as an aid makes a difference to the result. We can only state that the results without any aid are always poorer than the results when the testees could use one of the given tools.

The slightly poorer average of the third group, WN-FW-∅-MW, can possibly be explained by the the most difficult question set (21-30) being answered without any help.

Had we gotten more responses, we might be able to better distinguish between the different question groups and whether the order of the tools used matters. With the sample size being 34, we can only make careful guesses on what trend the results could follow.

		QUESTION GROUPS				
	N	1-10	11-20	21-30	31-40	MEAN (TOOLS)
∅-MW-PWN-FIWN	4	**8**	9	8.75	8.75	8.63
FIWN-∅-MW-PWN	7	9.14	**6.86**	8.43	9.43	8.47
PWN-FIWN-∅-MW	10	8.4	8.5	**5.7**	8	7.65
MW-PWN-FIWN-∅	13	8.85	9.08	8.54	**7.38**	8.46
MEAN (QUESTIONS)		8.6	8.36	7.86	8.39	
∅= NO HELP MW= MERRIAM-WEBSTER N= NUMBER OF ANSWERS						

Table 3: Results per question group and tool order

From this test set-up, however, we can draw conclusions on the usefulness of the tools. On the average, all testees got 6.99/10 questions correct without using any aids. The number tells us that the difficulty level of the questions is apt; in fact only 4 testees got a full score without using any help.

Using the Merriam-Webster dictionary improved most testees' performance. The number of perfect answers rises up to 10 when the testees get to use a dictionary as their aid.

Using the monolingual PWN as assistance, yields highest results. On the average the testees got 8.9/10 with the help of PWN. 12 of the answers were perfect. Based on the test, getting the Finnish translations alongside the English PWN does not improve performance on the sentence completion task. The number of perfect answers is 11 with the help of FiWN. We conclude that the translations do not provide additional value to the PWN in this type of a task.

3.1 The Wilcoxon Two Sample Test

We choose the Wilcoxon Two Sample Test[7] as the means for calculating statistical significance of the results. We want to see whether there is a significant difference in the way the testees performed while using different aids. The Wilcoxon Test fits our need because it does not assume the data to be normally distributed and yields accurate results with even small data.

We run the Wilcoxon Test on the material pairwise to see which tools differ from each other significantly. The test is performed for all possible pairings of the tools, as shown below. With the Wilcoxon test we can assume that if $p < 0.05$, it is not likely that the two groups have the same distribution and median making the difference statistically significant.

a) ∅ vs. Merriam-Webster

b) ∅ vs. PWN

c) ∅ vs. FiWN

d) Merriam-Webster vs. PWN

e) Merriam-Webster vs. FiWN

f) PWN vs. FiWN

We formulate the null hypothesis in the following way:
H_0: The data in groups x and y are independent samples from identical continuous distributions with equal medians.

We carry out the Wilcoxon tests to see whether we have to reject the null hypothesis at the 5% significance level. The results are given in Table 4.

TOOL	p <=
a) ∅ vs. M-W	0.00085
b) ∅ vs. PWN	0.000034
c) ∅ vs. FiWN	0.00033
d) M-W vs. PWN	0.3706
e) M-W vs. FiWN	0.7452
f) PWN vs. FiWN	0.5852

Table 4: Results of the Wilcoxon Two Sample test

The figures in Table 4 tell us that with the sample size of 34 at the 5% significance level we have to reject the null hypothesis for pairs d, e, and f. However, for pairs a, b and c, we cannot reject the null hypothesis. From this follows that we can

[7] http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html

assume that any of our chosen tools significantly helps the testee in completing the task.

Currently, the small number of responses prevents us from drawing firm conclusions on the significance of the difference between using wordnets and regular electronic dictionaries. However, the average performance using either of the two wordnets was better than using only an electronic dictionary.

4 Discussion and further work

The test gives us an insight into the usage of wordnets as dictionaries and into the way they can compete with traditional electronic dictionaries. The advantages a human user get from using a wordnet instead of a dictionary has so far not been widely studied.

We should extend the test with a larger sample of respondents to determine the significance of the improvement using wordnets over regular electronic dictionaries. Our number of responses is too small for making conclusions on which tool helps the testee most. Based on the experiment it is clear, however, that using any of our chosen tools helps the testee perform better.

Based on our study, it remains an open question whether the translations available in the FiWN bring any additional value to the testee. This could be better tested with a translation task, where the translated wordnet would probably be the most helpful tool. However, assessing the quality of the translations is difficult.

The average reported SAT results[8] in 2010 for the Critical Reading[9] part of the test for test takers with English as their first language are 64 percent.

Our sample consisted mainly of language students at the University of Helsinki and the sample performance of 69.9 percent on average conforms to the expectations when using no aids or tools. An initial concern that the performance boost is relevant only for non-native speakers therefore seems not to be the case.

Our experiment provided us feedback for development of the test and FiWN. After completing the test, the testees had the chance of leaving open feedback on both the testing method and the tools. Most testees found the translated wordnet most helpful, their gut feeling was that that using it would yield best results. Only a few testees preferred using Merriam-Webster over using either one of the wordnets.

Typically NLP applications that use wordnets for semantic classification use human performance as the gold standard when evaluating the results. This is the case e.g. in Turney and Littman (2005). The authors implement an algorithm for corpus-based learning of analogies and semantic relations and compare the results against human performance in the SAT analogy questions. There system correctly answers 47% of the questions where the average SAT test taker gets about 57% of the questions right.

We have established that human performance on tasks like sentence completion significantly improves if wordnets can be used as lexical aids. This most likely also applies to solving verbal analogies, since they are even more context-sensitive. We therefore suggest that NLP applications using wordnets should in fact be compared with human performance when humans use the same lexical resources.

To conclude, some studies have used wordnets to boost computer program performance on word sense disambiguation. Our study suggests that human users should perhaps be given a similar advantage if we wish to compare the results in a fair way.

5 Conclusion

Typically wordnets are used as lexical resources in various NLP applications. Using wordnets as lexical databases for other information systems has been studied widely, but the advantage wordnet provides to a human user as an electronic dictionary has received less interest.

We assessed the advantage of using a wordnet instead of a traditional dictionary as help in completing an SAT-type sentence completion task. The test was conducted as an online query divided in four parts. The purpose of the experiment was to see which lexical resources aid a non-native speaker the most. The resources we chose for the test are the Merriam-Webster online dictionary for the first set, the English WordNet for the second, and the bilingual FiWN, which can be used as an English-Finnish (and Finnish-English) dictionary for the third set of questions. To establish the En-

[8] http://professionals.collegeboard.com/profdownload/2010-total-group-profile-report-cbs.pdf

[9] The results for sentence completion are not given separately, so we have to compare our results to the Critical Reading section consisting of sentence completions and reading comprehension.

glish vocabulary skills, one set of questions was answered without any help.

The experiment sought to give insight on how useful wordnets are to a human user. The testees used both the English wordnet, and the bilingual FiWN so that we could test whether the translations bring any additional help to a non-native English speaker.

We found that a wordnet significantly improves the performance of a human user on a sentence completion task and we found weak indications that a wordnet may be slightly better than a regular electronic dictionary for this purpose. This sets new standards for what we should require from computers on similar tasks when comparing them with humans if we boost the computer performance with wordnets or other lexical resources.

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge/London/England.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.

Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors. 2007. *Proceedings of the Fourth Global WordNet Conference*. University of Szeged.

Peter D. Turney and Michael L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60:251–278.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Piek Vossen. 2004. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Linguistics*, 17(2):161–173.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955