Free construction of a free Swedish dictionary of synonyms

Viggo Kann and **Magnus Rosell**
KTH Nada
SE-100 44 Stockholm
Sweden
{viggo, rosell}@nada.kth.se

Abstract

Building a large dictionary of synonyms for a language is a very tedious task. Hence there exist very few synonym dictionaries for most languages, and those that exist are generally not freely available due to the amount of work that have been put into them.

The Lexin on-line dictionary[1] is a very popular web-site for translations of Swedish words to about ten different languages. By letting users on this site grade automatically generated possible synonym pairs a free dictionary of Swedish synonyms has been created. The lexicon reflects the users intuitive definition of synonymity and the amount of work put into the project is only as much as the participants want to.

Keywords: Synonyms, dictionary construction, multi-user collaboration, random indexing.

1 Introduction

The Internet has made it possible to create huge resources through voluntary cooperation of many people. The size of, or effort put into, each contribution does not matter – with many participators the sum may be great and useful. The most well-known example is the free-content encyclopedia Wikipedia[2] that anyone can edit. It has more than a thousand articles in 75 different languages. The English version has about 700,000 articles.

Wiktionary[3] is the lexical companion of Wikipedia. It is, as Wikipedia, a collaborative project, with the aim to produce a free dictionary in every language. The English Wiktionary has about 90,000 articles, and the Swedish about 3,750.

Both Wikipedia and Wiktionary use the copyleft[4] license GNU FDL[5], which means that the content is free to use. This often motivates people to contribute as they know that their work will be available to everyone.

To start a new similar project requires a lot of users that are interested in the matter and want to help. A suitable and very popular Swedish web site is the Lexin on-line dictionary[6]. Our plan was to let the Lexin users cooperate to build a free Swedish dictionary of synonyms.

To construct the dictionary of synonyms we followed these steps:

1. Construct lots of possible synonyms.
2. Sort out bad synonyms automatically.
3. Let the Lexin users grade the synonyms.

[1] http://lexin.nada.kth.se

[2] Wikipedia (http://www.wikipedia.org/) was founded in January 2001 by Jimmy Wales and Larry Sanger.

[3] http://www.wiktionary.org/

[4] http://www.gnu.org/copyleft/copyleft.html

[5] http://www.gnu.org/copyleft/fdl.html

[6] http://lexin.nada.kth.se

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 105–110
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

4. Analyze gradings and
   decide which pairs to keep.

The rest of the paper deals with these steps and presents the results so far of the efforts of the Lexin users. The project started in March 2005, and five months later a free Swedish dictionary of synonyms consisting of 60,000 graded pairs of synonyms was completed.

2 Possible synonym pairs

The first step is to create a list of possible pairs of Swedish synonyms. If you have access to a dictionary D_1 from Swedish to another language X and a dictionary D_2 from X to Swedish you can collect possible synonym pairs by translating each Swedish word to X and back again to Swedish, i.e.,

$$\{(w, v) : \exists y : y \in D_1(w) \wedge v \in D_2(y)\}$$

We may also consider only the dictionary D_1 from Swedish to X:

$$\{(w, v) : \exists y : y \in D_1(w) \wedge y \in D_1(v)\}$$

Similarly we may also consider only D_2. The pairs obtained in this way will sometimes be synonyms, but due to ambiguous word senses there will also be lots of rubbish.

If there are dictionaries available between Swedish and other languages one can get lists of word pairs from them using the same method. Such lists can then be used either to complement or to refine the original list. If (w, v) is a pair included in many lists it becomes more probable that w and v are real synonyms.

By using this technique we have constructed a list of 616,000 pairs of possible synonyms.

3 Automatic refinement

A possible way to improve the quality of the list would be to part-of-speech tag the words and only keep pairs containing words that may have the same word class. We chose not to do this, because words of different word classes could be (seldomly) synonyms, for example words of the word classes participles and adjectives. In retrospect it was a mistake not to remove words of different word classes, because it is annoying for many users to be asked whether for example a noun and a verb are synonymous.

We also refined the list of synonyms using a method called Random Indexing or RI (Kanerva et al., 2000). In RI each word is assigned a random label vector of a few thousand elements. Using these vectors one constructs a co-occurrence representative vector for each word by adding the random vectors for all words appearing in the context of each occurrence of the word in a large training corpus. For each word pair (w, v) the cosine distance between the co-occurrence vectors of w and v is a measure of relatedness between the two words; words that appear in similar contexts get a high value. Synonyms often appear in similar contexts. So this is a suitable method for deciding on whether a pair of possible synonyms are likely to be actual synonyms.

To find as many related words as possible we have used several different corpora. Table 1 gives some statistics for them. The three top sets are extracted from the KTH News Corpus (Hassel, 2001). Aftonbladet and DN are two Swedish newspapers and DI is a daily economic paper. Med is a set of medical papers from Läkartidningen[7] and Parole is a part of the Swedish Parole Corpus[8].

Before building the Random Index we removed stopwords and lemmatized all words. We chose random vectors with 1800 dimensions and eight randomly selected non-zero elements (four ones and four minus ones). When building the context vectors we used four words before and four

[7] http://www.lakartidningen.se/
[8] http://spraakbanken.gu.se/lb/parole/

Text set	Texts	Words	Lemmas
Aftonbladet	29,602	751,804	34,262
DN	6,954	593,055	63,164
DI	19,488	1,606,743	70,539
Med	2,422	2,146,788	150,627
Parole	–	1,694,556	135,205

Table 1: Text set statistics (stopwords not included)

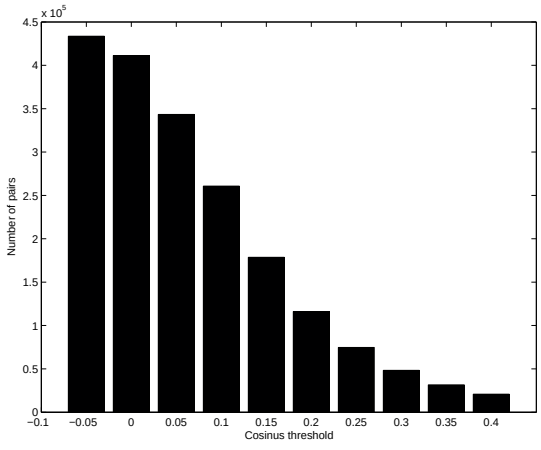

Figure 1: Number of pairs with cosine threshold

words after as the context for each word and added the random labels for these context words to the context vector of the center word weighted by 2^{1-d}, where d is the distance between the context word and the center word.

For each text set we then calculated the cosine of all word pairs in the list of possible synonyms and chose the maximum of these as their similarity. Of the 616,000 possible synonym pairs 435,000 appeared in any of the texts sets. Figure 1 shows the number of pairs (the vertical axis) with higher similarity than certain values (horizontal axis).

We chose to remove all pairs with a similarity value lower than 0.1 after studying this figure and some examples. This left 226,000 pairs.

4 Manual refinement

The Lexin on-line dictionary is a very popular web-site for translations of Swedish words to about ten different languages. During the year 2004 the number of lookups in Lexin was 101 millions. This means more than three lookups each second of the year. During 2005 this has increased to five lookups each second.

As the users of Lexin ask language (translation) questions they obviously like the idea of an on-line dictionary. Therefore they are probably motivated to put a small effort in producing a free Swedish dictionary of synonyms.

Many users are of course not native Swedes and are using Lexin to learn Swedish. In order to not bother them with questions about Swedish synonyms we chose to only include the synonym question in the Swedish-English dictionary with Swedish user interface. This will still cover two thirds of the total number of lookups of Lexin.

The Swedish Agency for School Improvement has allowed us to use the Lexin lookup answer web page. As a user gets an answer to a translation question she is also presented with the possibility to answer a question in order to help with the dictionary of synonyms. A question could for example be: *Are 'spread' and 'lengthen' synonyms? Answer using a scale from 0 to 5 where 0 means 'I do not agree' and 5 means 'I fully agree', or answer 'I do not know'*.

When a user have answered this question a web page of the growing synonym dictionary opens and the user may choose to grade more pairs, suggest new synonym pairs, lookup in the synonym dictionary or download the synonym dictionary. Prototypes of the programs taking care of the answers and the synonym dictionary were developed by a student project group at KTH.

5 Synonymity

It is interesting to note that the exact meaning of "synonym" does not need to be defined. The users will grade the synonymity using their intuitive understanding of the concept and the words in the question. The produced dictionary of synonyms will therefore use the People's definition of synonymity, and hopefully this is exactly what the people wants when looking up in the same dictionary.

Of this reason we called the dictionary *The People´s Dictionary of Synonyms*.

6 Abuse

Every web page that invites the public to participate will be subjected to attempts of abuse. Thus the synonym dictionary must have ways to prevent this. Our solution is threefold. First, many gradings of a pair are needed before it is considered to be a good synonym pair and become possible to lookup in the synonym dictionary.

Second, the pair that a user is asked to grade has been randomly picked from the list of about quarter of a million pairs. The same user will almost never be asked to grade the same pair more than once. If most of the users answer honestly and have an acceptable idea of the synonymity when they think they have, the quality of the synonym dictionary should be good.

Third, the word pairs that users suggest themselves are first checked using a spelling checker and are then added to the long list of pairs, and will eventually be graded by other users. The probability that a user will be asked to grade his own suggested pair is extremely small.

7 Results and discussion

In five months 2.1 million gradings were made (1.2 million gradings during the first two months). They were distributed over the different grades as shown in Figure 2. The distribution between grades 1–5 is remarkably even. Only the 0 grade is a lot more common than the other grades.

Table 2 gives a few examples of user graded synonyms. Pairs given grade 5 or 4 are very good synonyms. Pairs of grade 3 are synonyms to a less degree, for example *cistern* and *pot,* and pairs of grade 2 are often of different word classes (for example one adjective and one verb). Words of grade 1 are relatad but not synonymous, and grade 0 words are not even (obviously) related.

As the pairs are picked at random from the list some pairs are graded more times than others. Figure 3 shows the distribution of how many times the pairs were graded. Note that when a word has received three 0 gradings it will be removed from the list. Therefore there is a maximum point at 3.

In Figure 4 the number of pairs with different mean gradings are presented. Pairs with very small and very large mean gradings have several times during the period been removed from the list of words to be graded.

Figure 5 confirms that Random Indexing is useful for automatic refinement. The cosine similarity between pairs that are graded 0 by users is considerably lower than between all pairs on average.

The users could propose new synonym pairs. During the first five months 55,000 pairs (23,000 unique pairs) were proposed. After spelling correction end removal of non-serious words 15,000 pairs remained.

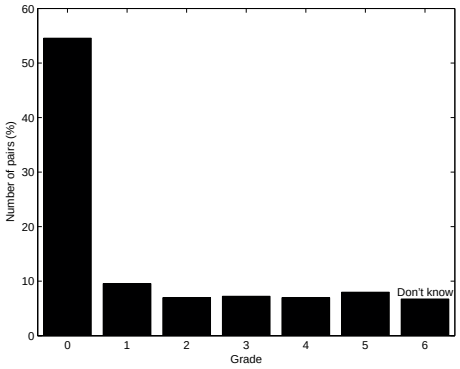

Figure 2: Gradings made by the users

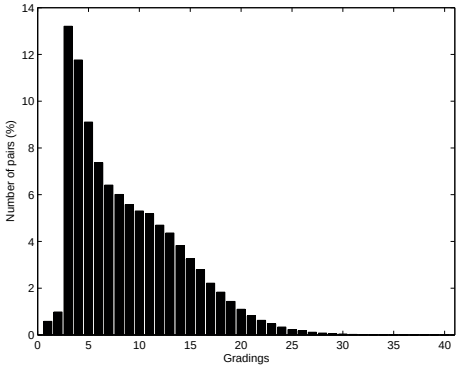

Figure 3: Number of gradings

These were regularly added to the list of pairs to be graded.

After five months we have 60,000 pairs in the dictionary of synonyms. All these pairs have been given a grade larger than 0 at least three times with a mean value of at least 2.0. When a user makes a lookup in the dictionary the synonyms are presented with their mean grades. This means that this dictionary of synonyms is much more useful than a standard one that only gives a bunch of words that may be more or less synonymous.

The 25,000 best synonym pairs have been publically available for downloading in a simple XML format for about a month. More than 50 downloadings each day are currently being performed, which shows that there is indeed a large need for a free dictionary of synonyms in Swedish.

Many users have commented that there were too many bad pairs. Lots of pairs were graded 0 (not at all synonyms) by all users. After some weeks 25,000 such pairs were removed. Later 60,000 more pairs were removed, improving the quality of the remaining pairs considerably.

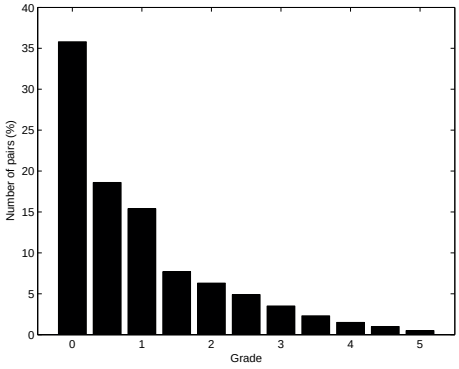

Figure 4: Mean gradings

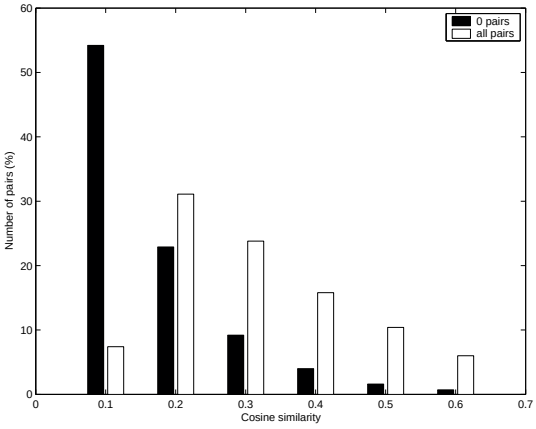

Figure 5: Cosine similarity for pairs graded 0 and for all pairs

8 Lessons learned

The list of suggested synonyms should be huge, as we want to find as many synonyms as possible. But bad pairs irritate the users.

5		4		3	
hipp	häftig	bisarr	konstig	cistern	burk
betraktare	iakttagare	dåraktig	idiotisk	folkskola	grundskola
gäng	grupp	hall	foajé	kamrat	väninna

2		1		0	
ansenlig	åtskilligt	bestiga	stiga	hake	kröka
fackförening	union	fatta	följa	ynklig	deltagande
glida	slinka	feja	ren		
hölja	omfatta	hård	tätt		

Table 2: Examples of user graded pairs

Therefore it is important to improve the quality of the list as much as possible. This could be done automatically, using for instance Random Indexing, word class tagging, and other dictionaries, for example for different languages. As the number of answers grows it is also a good idea to remove pairs that often get a zero grading.

9 Conclusions

We have found that it is possible to create a free dictionary of synonyms almost for free. The constructed dictionary is even more useful than a standard one, since the synonyms are presented with gradings.

There is no reason to believe that the method presented in this paper may not be used to create lists of other word relations, such as hypernymy, hyponymy, holonymy and meronymy.

The growing and improving dictionary of synonyms can be found at http://lexin.nada.kth.se/cgi-bin/synlex.

Acknowledgements

We would like to thank the Swedish Agency for School Improvement, and especially Kiros Fre Woldu, for letting us use the Lexin on-line web page to ask users for gradings of synonym pairs.

Thanks to the KTH students Sara Björklund, Sofie Eriksson, Patrik Glas, Erik Haglund, Anna Hilding, Nicholas Montgomerie-Neilson, Helena Nützman, and Carl Svärd for building the dictionary system prototype.

References

M. Hassel. 2001. Automatic construction of a Swedish news corpus. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA '01*.

P. Kanerva, J. Kristofersson, and A. Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proc. of the 22nd Annual Conference of the Cognitive Science Society*.