Semi-Automatic Analysis of Dictionary Glosses

Rune Lain Knudsen

Keywords: *wordnet, semantic networks, computational linguistics, bioinformatics.*

Abstract

Automatic methods for information retrieval, knowledge engineering/representation and text classification are important tools for processing large amounts of natural language. Lexicographic databases are being used as part of the toolset for some of these methodologies. At the Department of Linguistic and Scandinavian Studies (UiO), a Norwegian wordnet (NorNet) is being developed by applying a thorough analysis of the semantic parts of the definitions contained in Bokmålsordboka (BOB) in order to generate a network of semantic relations. In addition to the development of a wordnet using this dictionary-based method, the analysis stage of the process is valuable in itself as it can be used to give new insights into the consistency of the source material and gloss structure in general. An overview of the analysis stage is presented in this paper. The analysis is limited to verb definitions for the time being, and should be regarded as a work in progress.

1. Introduction

In my master's thesis I have explored a semi-automatic dictionary-based method for wordnet generation, which is inspired by Lars Nygaard's cand. philol. thesis *Frå ordbok til ordnett* (Nygaard, 2006). In Nygaard's thesis, an analysis was performed on a subset of the nouns represented in Bokmålsordboka (BOB), a dictionary for Norwegian Bokmål developed at the Department of Linguistics and Scandinavian Studies. The analysis generated a semantic network of hyperonym and synonym relations that served as the foundation for NorNet (Fjeld and Nygaard, 2009; Fjeld et al., 2012), a wordnet prototype for Norwegian Bokmål. My goal was to create a method was that will contribute to the extension of NorNet by generating relations for a different set of definitions and more classes of semantic relations. The main focus was on analyzing definitions for verbs. The method is explained in detail in Knudsen (2012), which will be publicly available this autumn.

The focus of the thesis was on the task of analyzing explanatory parts of verb definitions for the purpose of generating semantic networks. A thorough analysis of the dictionary was not the main subject of study. However, since the resulting semantic network is closely tied to the way the explanatory parts of definitions are formed, a number of observations made upon the source material can be done based on the analyses done by the method. In this paper I will give a brief overview over some of the parts that can be used to analyze definitions in a dictionary.

The method is currently restricted to an analysis of explanatory parts of verb definitions, which for the rest of this paper are referred to as *glosses*. Other grammatical classes, and other constituents of articles and definitions, are not considered as of now.

2. Preprocessing and extraction

Each gloss extracted from verb definitions in BOB is assigned a Part-of-Speech (PoS) sequence by OBT+Stat (Johannessen et al., 2011), a morphological and syntactic tagger developed by Tekstlaboratoriet, UiO, and Uni Computing. Table 1 shows some examples of tagged glosses. A set of PoS pattern classes are defined, each one being defined by a distinct PoS tag sequence. Each gloss is assigned to the PoS pattern class whose PoS tag sequence matches that of the gloss. The size of a PoS pattern class is measured in terms of its number of

members. In total, 9927 glosses for verb definitions have been extracted from BOB. After running the morphological tagger on these glosses, 4007 distinct PoS pattern classes are generated.

Table 1. Some examples of glosses tagged by OBT+Stat. English translations are quoted. All English words should be considered verbs.

Definiendum	**Gloss**			
terminere	begrense	,	avslutte	
'terminate'	'limit'	,	'end'	
	VERB	KOMMA	VERB	
referere	henvise			
'refer'	'direct'			
	VERB			
undertrykke	hemme	,	holde	tilbake
'suppress'	'inhibit'		'hold'	'back'
	VERB	KOMMA	VERB	PREP
etse	la	tære	,	oppløse
'etch'	'let'	'corrode'	,	'dissolve'
	VERB	VERB	KOMMA	VERB
besvime	miste	bevisstheten	,	dåne
'faint'	'loose'	'consciousness'	,	'swoon'
	VERB	SUBST	KOMMA	VERB

The largest PoS pattern class is one identified by two verbs separated by a comma. This POS pattern class contains 809 glosses, which accounts for approximately 20:2% of all the extracted verb glosses. The second most frequent pattern consists of a single verb, and contains 650 glosses. This accounts for approximately 16:2% of all the extracted verb glosses. There are 3395 PoS pattern classes of size 1, which means that approximately 34:3% of all extracted verb glosses are distinct in terms of morphosyntactic structure. By performing a similar analysis on other dictionaries, comparisons can be done and claims be made regarding strategies for text condensing of glosses and the consistency of the application of said strategies.

3. Manual transducer generation

The set of PoS pattern classes is the foundation for further analysis of the glosses. To generate semantic relations such as hyperonyms, synonyms, causal relations et cetera, *transducers* are applied to the PoS patterns. Transducers belong to a set of algorithms that transform some input sequence of tokens into another output sequence based on some ruleset. In this case, the input is a sequence of PoS tags, and the output is a sequence of semantic relations. The ruleset is defined by regular expressions (Jurafsky and Martin, 2008, p. 51-77), a formal language that specifies search text strings by using a set of operators specifying symbol grouping (denoted by parentheses) repetitions (denoted by an asterisk), and many more.

3.1. *Synonym patterns*

For example, to generate synonym relations for single-verb glosses and comma-separated verbs, a transducer with an input rule VERB (KOMMA VERB)* (i.e. a VERB token followed by zero or more KOMMA VERB sequences) and an output SYN (NIL SYN)* (i.e.

HAS_SYNONYM relations for all subsequent verbs and no relation for commas) could be defined. Some examples of such a transducer and resulting relations can be seen in Table 2.

Table 2. Examples of glosses matched by the transducer specified in the first two rows, along with generated relations. English translations are quoted.

Input	VERB	(KOMMA	VERB)*		
Output	SYN	(KOMMA	SYN)*		
verge	forsvare	,	verne		
'guard'	'defend'	,	'protect'		
uttrykke	formulere	,	uttale		
'express'	'formulate'	','	'pronounce'		
avsløre	avduke				
'reveal'	'unveil'				
assistere	hjelpe	,	bistå	,	medvirke
'assist'	'help'	','	'assist'	','	'contribute'
verge	HAS_SYNONYM	**forsvare**			
verge	HAS_SYNONYM	**verne**			
uttrykke	HAS_SYNONYM	**formulere**			
uttrykke	HAS_SYNONYM	**uttale**			
avsløre	HAS_SYNONYM	**avduke**			
assistere	HAS_SYNONYM	**hjelpe**			
assistere	HAS_SYNONYM	**bistå**			
assistere	HAS_SYNONYM	**medvirke**			

3.2. *Hyperonym patterns*

Other interesting aspects of more complex verb definitions can also be investigated. One example is the question of whether or not verb definitions can be formed according to Aristotelian principles. This is often the case for noun definitions, but verb definitions are not necessarily as obvious.

Consider the gloss for the verb **slipe 'grind'**: *kvesse redskap med egg 'sharpen tools with edges'*. The gloss assigns **slipe** to the class of **kvesse**, distinguishing it from other uses of **kvesse** by specifying it to an action done for certain tools - in this case tools with edges like knives, axes and so forth. This definition belongs to the PoS pattern class "VERB SUBST PREP SUBST".

When looking at other members of this PoS class, we find a number of glosses conforming more or less to the same principle. **orkestrere 'orchestrate'** has a gloss *arrangere musikk for orkester 'arrange music for orchestra'*. **frede 'preserve'** has a gloss *verne dyr ved fangstforbud 'protect animals by hunting ban'*. We can however spot a tendency that seems to apply to a lot of verb definitions: it is not always clear whether a verb should be considered a hyperonym of another verb, or a synonym.

Another frequent pattern was the case of one verb, followed by a comma, followed by a short explanatory sentence. In many cases, this pattern seemed to follow the same principle as the one above, in that the first verb in the gloss could be interpreted to be the closest hyperonym. Some inconsistency was nonetheless observed. For example, consider the glosses for **fratre 'retire'** and **adherere 'adhere'** in Table 3. In the case of **fratre**, the first verb in the gloss has little relationship to the rest of the gloss, and can not substitute the next verb (**trekke**) without damaging the overall syntactic coherence of the gloss. The first verb for **adherere** however, has a closer relationship to the rest of the gloss in that the two verbs could easily replace one another. Such an observation might be an indication of an inconsistency in that particular area of the dictionary.

Table 3. Examples of possibly inconsistent glosses according to PoS pattern similarity. English translations are quoted.

Lemma	Gloss				
anføre	nevne	,	påberope	seg	
'note as'	'mention'	','	'claim'	'oneself'	
fratre	slutte	,	trekke	seg	tilbake
'retire'	'quit'	','	'withdraw'	'oneself'	'back'
adherere	henge	,	holde	fast	ved
'adhere'	'cling'	','	'hold'	'on'	'to'

Another phenomenon that occurs within many verb definitions is that glosses conforming to the genus proximum / differentia specifica tend to assign a very general hyperonym to the definiendum. Looking at the same PoS pattern class, **synge 'sing'** has a gloss *frambringe toner med stemmen 'produce tones with the voice'*. **evaporere 'evaporate'** has a gloss *lage ferskvann ved hjelp av evaporator 'create fresh water by means of an evaporator'*. **fri 'propose'** has a gloss *gi tilbud om ekteskap 'give an offer of marriage'*. General verbs such as **gi 'give', gjøre 'do', bli 'become'** and **lage 'create'** thus represent a substantial amount of candidate hyperonyms. In many cases this is to be expected as hyperonym/troponym hierarchies for verbs are known to be shallow (Fellbaum, 1998, p. 80), especially when compared to noun hyperonym/hyponym hierarchies. There is nonetheless a motivation for examining such definitions in a dictionary in order to ensure that the use of these general verbs do not occur more often than necessary.

4. Operator word assignment

Transducers provide an efficient way to analyze simple verb glosses like the ones shown in Table 2, but for more complex glosses we need to add some flexibility to the algorithm. Since troponym hierarchies tend to be shallow(Fellbaum, 1998, p. 80), the need for additional relation types further specifying the meanings of concepts is perhaps even more important for verbs than for nouns. As an example, consider the gloss *forsterke, gjøre mer effektiv 'strengthen, make more effective'* for the definiendum **intensivere 'intensify'**. We desire to generate relations between the definiendum and the words **forsterke** and **effektiv** at minimum. It is clear that gjøre affects the semantic value of **effektiv** by specializing it, implicating that the definiendum has a causal relationship to something, effektiv being part of the outcome in some way. If we ignore **gjøre**, we no longer know if effektiv is an inherent property of that something, or the effect of some cause.

To account for such phenomena, functionality for transforming a sequence of semantic relations based on lemma information is implemented. From this point and onwards, we make a distinction between *target words* and *operator words*. Target words refer to tokens in a gloss that contain semantic information about its definiendum. Operator words refer to tokens in a gloss that do not have much semantic content themselves, but that instead change the meanings of tokens in their surrounding context. Looking at our example, the words **forsterke** and **effektiv** are considered target words according to our distinction, while **gjøre** is considered an operator word. By turning operator words into actual operators that explicitly change the semantics of tokens surrounding it, we can automatically transform the output of the transducers in order to account for this behavior.

Table 4. Some examples where the use of operator words are needed. Definienda are marked with boldface. English translations are quoted.

	VERB	ADJ
svekke	gjøre	svak
'weaken'	'make'	'weak'
kjærtegne	berøre	ømt
'caress'	'touch'	'tenderly'
kjølne	bli	kaldere
'cool'	'become'	'colder'

Candidate operator words are found by making a frequency list over all lemma forms of the tokens found in glosses. The most frequent lemma forms tend to belong to the grammatical class of prepositions, adjectives and adverbs, very general concepts close to the top of a troponym hierarchy (e.g. **person 'person', land 'country'**), and a number of lemmas that belong to the class of words that is commonly referred to as *light verbs* (Butt, 2003**). Være 'be', gjøre 'do/make'** and **bli 'become'** are some examples. Table 4 shows some instances where for example **gjøre** influences the following adjective. **Svekke** has **svak** as an adjective in its gloss, which could be either a sort of implication, manner, or an effect of the definiendum. But if we have defined **gjøre** as an operator word that changes the semantic state of a following adjective into an effect, we can infer with a high confidence that **svekke** indeed has a CAUSE relationship to **svak**. The same goes for **bli 'become'**, resulting in a relationship *kjølne CAUSES kaldere*.

5. Semi-automatic transducer expansion

The proposed method is based on the assumption that similar glosses exhibit similar behaviour. A technique for comparing POS pattern classes that share certain properties is thus needed. This is accounted for by employing a local alignment algorithm, specifically the Smith-Waterman algorithm (Smith and Waterman, 1981).

The Smith-Waterman algorithm belongs to a class of dynamic programming algorithms that align two sequences in order to identify similar sub-sequences and to give a measure of this similarity. It has been applied to of a number of problems both in the field of bioinformatics (Smith and Waterman, 1981) and natural language processing (Katrenko et al., 2010). In this case, the algorithm is used to compare our PoS pattern classes, and to align both the tagged sequence and its corresponding sequence of relations generated by a transducer. A similarity metric based on the Jensen-Shannon divergence (Lin, 1991) is applied to the glosses to account for the co-occurrences of operator words, further refining the similarity score. When ranking the resulting POS pattern alignments according to the alignment score and the Jensen-Shannon divergence, a selection of the most similar POS patterns can subsequently be analyzed further through the steps previously described. An example of this is given in Table 5.

Table 5: A selection of some of the high-scoring alignments resulting when comparing the PoS pattern VERB ADJ with other PoS patterns. The definiendum is shown in boldface. English translations are quoted.

	VERB	ADJ	—-
	VERB	ADJ	PREP
gjennomlyse	lyse	helt	igjennom
'illuminate'	'shine'	'all the way'	'through'
	VERB	—	ADJ
	VERB	ADJ	ADJ
dovne	bli	midlertidig	følelsesløs
'(to) numb'	'become'	'temporarily'	'numb'
vakne	bli	fullt	bevisst
'wake (up)'	'become'	'fully'	'conscious'
avdramatisere	gjøre	mindre	dramatisk
'downplay'	'make'	'less'	'dramatic'
	VERB	ADJ	—-
	VERB	ADJ	SUBST
formulere	gi	språklig	form
'formulate'	'give'	'linguistic'	'form'

By applying transducer expansion, one can examine additional glosses similar to the ones initially chosen without specifying additional search patterns. This will show, among other things, if similar glosses do in fact exhibit similar properties. As an example, consider the gloss for **dovne**. The adjective in the original sequence, along with its corresponding relation, is shifted one position to the right. This causes the algorithm to ignore the adjective in the middle (**midlertidig**), and thus generate the appropriate *dovne CAUSES følelsesløs* relation. As can be seen in the gloss for **avdramatisere**, an erroneous relation *avdramatisere CAUSES dramatisk* is created. This can however be remedied by assigning some operator word functionality to **mindre 'less'**.

6. Conclusion

This paper has presented some examples of how a semi-automatic dictionary-based method for wordnet generation can provide insights into the explanatory part of verb definitions. Some examples have been presented of how Part-of-Speech tagging and algorithms from bioinformatics and natural language processing can present dictionary data in a way that lets us explore the dictionary from a different angle, be it by means of other digital interfaces or by book form.

As this is a work in progress, the various parts of the method still need to be refined in order to achieve the best possible consistency with regards to the task of automatically generating a semantic network of high quality. The refinement process, and observations done during this process, should nonetheless be of value as long as one can systematically discover errors and trace them back to transducers, certain operator words, or the dictionary itself. The examples shown are fairly simple, but the method could easily be extended to other parts of definitions, other grammatical classes, other relations, et cetera. One might discover ways to improve dictionary definitions that otherwise would be left unnoticed for a long time, shared patterns recurring in definitions, patterns - or lack of patterns - in the use of certain keywords, light verbs, prepositions, placeholders, and so on.

References

Butt, M. 2003. 'The Light Verb Jungle.' *Harvard Working Papers in Linguistics* 9: 1–49.

Fellbaum, C. 1998. 'Semantic network of English verbs.' In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press, 69–104.

Fjeld, R. V. and L. Nygaard 2009. 'NorNet - a Monolingual Wordnet of Modern Norwegian.' In B. S. Pedersen, A. Braasch, S. Nimb and R. V. Fjeld (eds.), *NEALT Proceedings Series, volume 7*. Northern European Association for Language Technology, 13–16.

Fjeld, R. V., R. L. Knudsen and J. M. Torjusen Forthcoming. 'Fra alfabet til begrep: Bokmålsordboka og NorNet.' In *Nordiska Studier i Lexikografi 11. Rapport från Konferensen om leksikografi i Norden, Lund 24 - 27 mai 2011*.

Johannessen, J. B., K. Hagen, A. Nøklestad and A. Lynum 2011. 'OBT+Stat: Evaluation of a Combined CG and Statistical Tagger.' In E. Bick, K. Hagen, K. Müürisep and T. Trosterud (eds.), *NEALT Proceedings Series, volume 14*. Northern European Association for Language Technology, 26–34.

Jurafsky, D. and J. H. Martin 2008. *Speech and Language Processing*. (2nd Edition) Prentice Hall. (Prentice Hall Series in Artificial Intelligence).

Katrenko, S., P. Adriaans and M. van Someren 2010. 'Using local alignments for relation recognition.' *Journal of Artificial Intelligence Research* 38:1–48.

Knudsen, R. L 2012. *Wordnet semantics from dictionaries*. MA Dissertation, University of Oslo.

Lin, J. 1991. 'Divergence measures based on the shannon entropy.' *IEEE Transactions on Information Theory*. IEEE Information Theory Society, 145–151.

Nygaard, L. 2006. *Frå ordbok til ordnett*. Cand. Philol. Thesis, University of Oslo.

Smith, T. F and M. S. Waterman 1981. 'Identification of common molecular subsequences.' *Journal of Molecular Biology* 147.