Semantic Mirrors

Helge Dyvik[1]

Abstract. We motivate and present a method for deriving lexical semantic information from translational data. Based on the network of translational relations between the words of a word-aligned and lemmatized parallel corpus the method individuates word senses, places them in structured semantic fields, derives lattices showing semantic relationships, and derives parametrized thesaurus-like entries from the lattices. The method has been applied to the English-Norwegian Parallel Corpus ENPC.

1 INTRODUCTION

There is a growing body of work based on the assumption that translations can be exploited as sources of information about semantics, e.g. (Resnik and Yarowsky 1997), (Ide 1999a and 1999b), (Diab and Resnik 2002), (Ide & al. 2002), (Tufis and Ion 2003), (Tufis & al. 2003), (Pianta and Bentivogli 2003), (Tufis & al. 2004), (Priss and Old 2005).

An approach which avoids the reliance on existing semantic resources or annotation is attempted in (Dyvik 1998a, 1998b, 2002, and 2005). The approach, dubbed 'Semantic Mirrors', takes as its point of departure only information about the set of possible translations of each of a set of lemmas in a word-aligned and lemmatized parallel corpus.[1] The aim is to achieve (i) the individuation of contrastively different senses for each lemma, (ii) grouping of the individuated senses across lemmas in structured semantic fields, (iii) derivation of feature-based representations of each sense in a field, allowing the construction of a semilattice revealing semantic distance as well as hyperonym/hyponym relations among the senses, and (iv) derivation of thesaurus-like entries for lemmas from the semilattices, in which the 'steepness' of the hyperonym/hyponym hierarchies and the granularity of the further division of senses into related subsenses is parametrized.

2 ASSUMPTIONS

The Semantic Mirrors method takes the translational relation between two languages as a theoretical primitive, making the following assumptions:

1. Semantically closely related words tend to have strongly overlapping sets of translations.
2. Words with wide meanings tend to have a higher number of translations than words with narrow meanings.
3. If a word sense *a* is a hyponym of a word sense *b* (such as *tasty* of *good*, for example), then the possible translations of *a* will probably be a subset of the possible translations of *b*.[2]
4. Contrastive ambiguity, i.e., ambiguity between two unrelated senses of a lemma, such as the two senses of *band* ('orchestra' and 'piece of tape'), tends to be a historically accidental and idiosyncratic property of individual lemmas. Hence we do not expect to find instances of the same contrastive ambiguity replicated by other lemmas in the language or by lemmas in other languages.[3]
5. Words with unrelated meanings will not share translations into another language, except in cases where the shared target lemma is contrastively ambiguous between the two unrelated meanings. Crucially, by assumption 4 there should then be at most one such shared lemma.

3 DEFINITIONS

In the following 'word' means 'lemma' in the sense of footnote 2. By the 'translations' in a language L2 of a word *w* in a language L1 we will understand both the words *into* which it is translated in L2 and the L2 words *from* which it is translated when L1 is the target language. Thus we disregard the direction of translation and consider only the symmetrical relation of *translational correspondence*.[4]

3.1 *t*-images and Sense Individuation

The first t-image in L2 of a word *w* in L1 is the set of translations of *w* in L2. *The inverse t-image* of *w* is the set of first *t*-images in L1 of each of the members of the first *t*-image of *w*. *The second t-image* of *w* is the set of *t*-images in L2 of each of the members of the union of the inverse *t*-image.

Figure 1 shows the first and inverse *t*-images of a word *sa*, where the arrow can be read 'has as its set of alternative

[1] Department of Linguistic, Literary and Aesthetic Studies, University of Bergen, N-5007 Bergen, Norway. Email: helge.dyvik@lle.uib.no

[1] Since the identification of homonymy is an intended outcome of the method, we do not presuppose that homonyms have been separated in the input. Therefore, a 'lemma' may correspond to a set of homonymous lexical entries.

[2] The likelihood of finding the de-specifying or specifying translations presupposed here increases with the vagueness of the words involved.

[3] Borrowings provide counterexamples, but we assume only with a marginal effect.

[4] This does not imply the empirical assumption that the translational relation is symmetrical in the sense that whenever *a* can be translated with *b*, *b* can also be translated with *a*. We are simply considering the union of translational correspondences, disregarding direction.

translational correspondents'. The putative individuation of contrastively different senses (i.e., the identification of homonymy) is based on the *t*-images. *sa* will necessarily be a member of all the sets in its inverse *t*-image. By assumptions 4 and 5, if the intersection of the *t*-images of two words t_i and t_j contain at least one member in addition to *sa*, then t_i and t_j are taken to be semantically related and therefore to reflect the same sense of *sa*. On the other hand, if the intersection contains only *sa*, then t_i and t_j are taken to be semantically unrelated and therefore to reflect contrastively different senses of *sa*. The result is a partitioning of the first *t*-image of *sa* into sense partitions, where each partition contains semantically related words,[5] which in the case of Figure 1 yields: {{*ta, tb*}, {*tc*}, {*td, te*}}. On the basis of the sense partitions three senses of *sa* are distinguished: *sa-1*, *sa-2* and *sa-3*. Each sense takes one sense partition as its first *t*-image, *sa-1* being associated with {*ta, tb*}, etc. Once this procedure has been applied to all the words of L1 and L2, the words in the sense partitions can be replaced by their relevant senses. At that point the first *t*-image of the L1 sense *sa-1* may be, say, the set of L2 senses {*ta-3, tb-1*}.

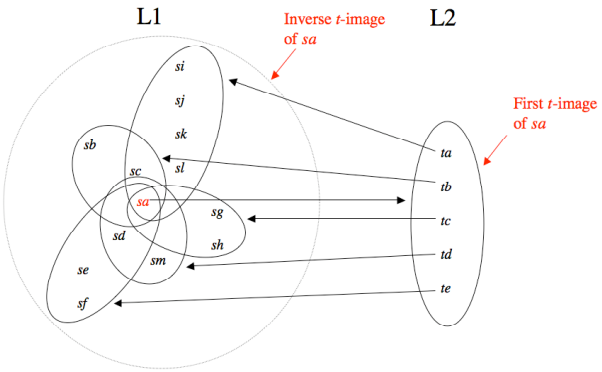

Figure 1. The first and inverse *t*-images of a word *sa*.

3.2 Semantic Fields

Next, the word senses across different lemmas in each language are grouped into *semantic fields* based on shared translational properties. Two L1 senses *a* and *b* belong to the same semantic field iff either (i) they have intersecting first *t*-images, or (ii) there is a sequence of such intersecting *t*-images joining them.

The semantic fields in L1 and L2 will necessarily be paired one-to-one, given the symmetry of the translational relation which determines field membership. Each field in such a pair projects a subset structure onto the other field, since the first *t*-image of each member of a field *F1* in L1 is a subset of the corresponding field *F2* in L2. According to assumptions 1-3 this subset structure contains rich information about the semantic relationships among the field members. Thus, according to assumption 2, an L1 sense which is a member of many subsets and hence has many translational partners in the L2 field, will have a wider meaning than a sense which is a member of few subsets. Furthermore, according to assumption 3, if the sets of which a sense *a* is a member constitute a subset of the sets of which a sense *b* is a member, then *a* is expected to be a hyponym of *b*.

Figure 2 shows a structured semantic field[6] which has been derived from the English-Norwegian Parallel Corpus (ENPC).[7]

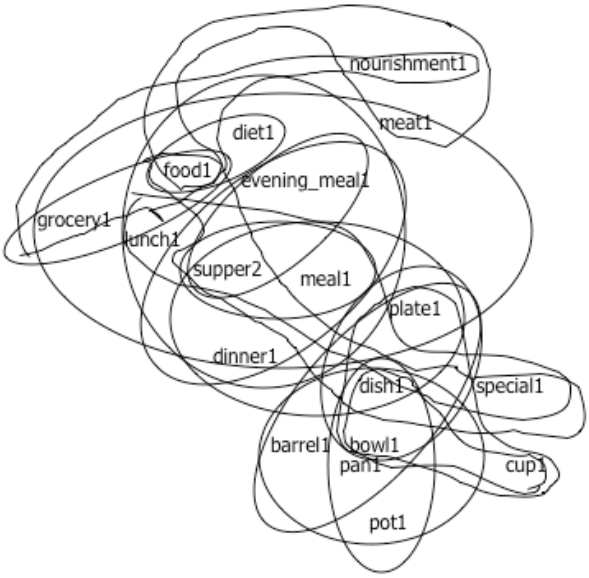

Figure 2. A structured semantic field derived from ENPC.

The field in Figure 2 comprises both food-related and vessel-related noun senses. They have been joined in one semantic field through the semantic ambivalence of the senses *dish1* and *plate1*, both of which can denote portions of food as well as food vessels, as reflected in their Norwegian translations (not given here).

The field in Figure 2 also indicates that senses like *food1* and *supper2* are wider than, e.g., *grocery1* and *meat1*, being members of a higher number of subsets.

3.3 Semantic Features

Based on a pair of structured semantic fields in L1 and L2 we proceed to assign a *semantic representation* to each sense in each field, encoding its relationships to the other senses in its field. The representations take the form of *feature sets*, and the aim is to assign few features to wide senses and supersets of those features to their hyponyms. A hyponym is allowed to inherit features from more than one hyperonym hierarchy. Features are constructed from translationally paired L1 and L2 senses and assigned to the senses from which they are constructed as well as to other senses which may inherit them.

[5] Sparseness of corpus data will often lead to too many putative senses being distinguished in this way. In practice one can therefore regard the sense partitions as a *partial* clustering of sense-related translational correspondents.

[6] I am grateful to Gunn Inger Lyse (2003) for the data on which this example is based.

[7] The ENPC contains fiction as well as non-fiction texts and translations in both directions. It comprises approximately 2.6 million words. See (Johansson & al. 1996), (Johansson & al. 1999/2002).

We refer to the number of *t*-image subsets of which a sense is a member as its 'rank'. Feature assignment to the senses of the paired fields *F1* in L1 and *F2* in L2 starts from the peaks, where the 'peak' of a field is the sense with the highest rank.The peak of the field in Figure 2 is *supper2*. A feature [a|b] constructed from source sense *a* and target sense *b* is assigned to *a* and *b* and inherited by all members of the first *t*-images of *a* and *b* which are ranked lower than *b* and *a*, respectively.

By hypothesis, feature set inclusion now expresses a hyponymy/hyperonymy relation. Thus, after feature assignment to the field in Figure 2, the senses *food1* and *lunch1* have the features shown in Figure 3, where the features of *food1* is a subset of the features of its hyponym *lunch1*:

food1	**lunch1**
[mat1|supper2]	[mat1|supper2]
[middag1|food1]	[middag1|food1]
	[lunsj1|meal1]
	[lunch1]

Figure 3. Feature assignment to two senses.

The full set of senses in a field is thus partially ordered by set inclusion. We can construct an upper semilattice from this set, allowing us to compare the distances between all the senses in the field. An upper semilattice in our case means that for each pair of feature sets, either one set includes the other or, if they intersect, there must be a third feature set consisting of the intersection of the two sets. By adding elements with such intersections whenever they do not exist already, we construct an upper semilattice from a semantic field. We label the added elements as indexed *x*'es.

Figure 4 shows a sublattice of the lattice constructed from the semantic field in Figure 2.[8] In Figure 4 only the senses *dish1* and *plate1* are dominated both by nodes dominating vessel senses and by nodes dominating food senses, which indicates their status as belonging in two hyperonymy hierarchies.

3.4 Thesaurus Derivation

Thesaurus-like entries can be derived from the lattices by abstracting away from some of the (probably partly spurious) detail in them. The derivation can be parametrized in order to allow alternative strategies for reducing the information in the lattices. We parametrize two aspects of the thesaurus entries: the steepness of the hyperonymy/-hyponymy hierarchy and the granularity of the further division of senses into related subsenses.

In order to maintain a plausible concept of 'hyperonym' we may want to set a lower bound on its number of hyponyms. A sense dominating only two or three other senses in a large lattice may more plausibly be considered as their near-synonym than as their hyperonym. We do this by means of a variable *SynsetLimit* which specifies the number of senses that must have inherited a feature *f* constructed from a sense *s* for *s* to be counted as their hyperonym. *SynsetLimit* can be set manually or set to vary as a function of the size of the semantic field.

Furthermore, the sense *s* can be divided into mutually related subsenses. Each feature assigned to *s* potentially represents a distinct subsense; whether two features *f1* and *f2* should be considered as belonging to the same subsense or not can be determined on the basis of the sets of senses to which *f1* and *f2* are assigned. If the cardinality of the intersection of these sets of senses exceeds a certain parametrized threshold called *OverlapThreshold*, then the features are not considered as representing distinct subsenses. The *OverlapThreshold* has a value between 0 and 1, representing a degree of overlap between two sets. Hence, in general, the higher the *OverlapThreshold* is set, the more subsenses are distinguished, with one subsense per feature as the theoretical maximum.[9]

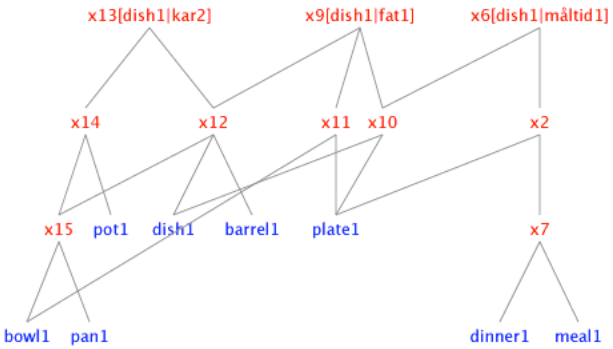

Figure 4. A sublattice.

Example (1) shows the derived thesaurus entry for the relevant sense of *dish* with *SynsetLimit* = 10 and *OverlapThreshold* = 0.1, while (2) shows the result of increasing *OverlapThreshold* to 0.2. The increase leads to the separation of the vessel and food senses. The ambivalent sense of *plate* makes it occur as a synonym of both subsenses.

(1) *OverlapThreshold* = 0.1:

dish
(*Translation:* måltid, kar, fat)
Synonyms: barrel, bowl, dinner, meal, pan, plate, pot.

(2) *OverlapThreshold* = 0.2:

dish
Subsense (i)
(*Translation:* måltid)
Synonyms: dinner, meal, plate.
Subsense (ii)
(*Translation:* fat, kar)
Synonyms: barrel, bowl, pan, plate, pot.

[8] The top single-feature *x*-nodes in Figure 4 are displayed with their feature contents. Obviously the full lattice contains other nodes dominating nodes in this sublattice.

[9] The derivation of thesaurus entries from translational data, and the effect of changing the parameters, can be tested here: http://decentius.aksis.uib.no:83/~helge/mirrwebguide.html

Decreasing *SynsetLimit* to 3 changes the status of the synonyms of subsense (ii) in example (2) to hyponyms.

Examples (4)-(5) show the effect on the entry for *authentic*[10] of increasing the value of *SynsetLimit*: hyperonyms are redefined as synonyms, and further synonyms and related words appear. The latter are senses which share a feature which previously defined a hyperonym. Thus an increased *SynsetLimit* reduces the 'height' of the hyperonym/hyponym hierarchies and increases the 'breadth' of synsets by joining them; the fields thus become less 'steep'.

(4) SynsetLimit = 5:

authentic
(*Translation:* virkelig, oppriktig, ekte, egentlig)
Hyperonyms: honest‹1› , true‹1› , right‹2› .
Synonyms: genuine‹1› .
Related words: legitimate‹1› , sincere.

(5) SynsetLimit = 15:

authentic
(*Translation:* virkelig, oppriktig, ekte, egentlig)
Hyperonyms: true‹1› .
Synonyms: genuine‹1› , honest‹1› , legitimate‹1› , right‹2› , sincere, truthful‹1› .
Related words: accurate, actual‹1› , confident‹2› , even‹3› , frank‹1› , proper‹1› , regular‹1› , serious‹1› , smart‹1› .

4 CORPUS DATA AND EVALUATION

The Semantic Mirrors method has been applied to data from the English-Norwegian Parallel Corpus (ENPC).[11] The examples discussed so far in this paper have been based on manual word alignment of all the relevant corpus occurrences. In addition automatic word alignment, developed by Sindre Sørensen at Aksis, Bergen, was applied to the ENPC, allowing us to extract the first *t*-images for 21153 Norwegian and 13344 English nouns, 3043 Norwegian and 2983 English verbs, and 4308 Norwegian and 4003 English adjectives. (9) gives a minuscule glimpse of the English thesaurus derived from this automatically provided material (*SynsetLimit*: automatic by size of the semantic field, *OverlapThreshold* = 0.05).

(9) A glimpse of a thesaurus based on automatically word-aligned data.

creature (n)
Sense 1
Subsense (i)
(*Translation:* skapning)
Synonyms: animal‹2› , organism‹2› .
Subsense (ii)
(*Translation:* vesen)
Synonyms: nature‹2› .
Related words: being‹3› , character‹2› , manner‹1› , personality‹2› , presence‹2› , result‹2› , someone‹1› .
Sense 3
(*Translation:* utyske)
Synonyms: monster‹2› , ogre.
credit (n)
(*Translation:* fordel, favør)
Synonyms: benefit‹3› , favor, favour‹1› .
Related words: advantage‹2› , expense‹1› , government‹3› .
credit (v)
(*Translation:* tillegge)
Hyperonyms: have‹1› .
Synonyms: shall.
Related words: add‹1› , attribute‹1› , invest‹1› .
creek (n)
(*Translation:* bekk)
Synonyms: rivulet.
creep (v)
(*Translation:* liste, snike)
Synonyms: tiptoe‹1› , cast, slink, sneak.
Related words: list.
creepy (a)
(*Translation:* skummel)
Synonyms: fishy.

The word aligner has an estimated precision of 84% and an estimated recall of 62%.[12] Comparison of the results from automatic alignment with those from manual alignment clearly shows that the method is vulnerable to noise in the word alignment. Thunes (2003) considers a selection of 43 adjectives and shows that if we take as a gold standard the joined sets of the synonyms, related words, hyperonyms and hyponyms of each adjective established on the basis of *manual* word alignment, and compare with the corresponding sets for the same adjectives established on the basis of *automatic* word alignment, then the precision of the latter is 35% and the recall is 14%. This is significantly lower than the estimated precision and recall of the word alignment itself (84% and 62%, respectively). In other words, small errors in the word alignment are magnified by the Mirrors method.

As is well known, the notion of an uncontroversial universal gold standard in the field of semantic resources is problematic. Therefore comparisons taking existing resources as gold standards can hardly be purely quantitative – the results must be qualitatively evaluated. Thunes (2003) performs an evaluation of the entries for 43 adjectives in an ENPC-based Mirrors thesaurus based on manual word alignment, taking the corresponding entries in Merriam-Webster's Thesaurus as a gold standard. The evaluation disregards the distinction between hyperonyms, hyponyms, synonyms and related words, taking them jointly as a set of 'R-words' of a given lemma. The R-word sets are then compared to the corresponding R-word sets in Merriam-Webster. Thunes provides an indication of the informativeness of her quantitative results by using the Princeton Wordnet as an alternative gold standard for a subset of the adjectives. Figure 5 shows the R-word sets for the adjective *pleasant* in the three resources:

[10] This example is also based on manually word-aligned material from the ENPC. The semantic field in this case contains hundreds of senses.

[11] Apidianaki (2008) applies it to a selection of data from an English-Greek parallel corpus.

[12] The estimation is based on two randomly selected sentence pairs from each of the 100 text pairs in the corpus.

Figure 5. 'R-words' for the adjective *pleasant* in three resources, after (Thunes 2003).

The Mirrors precision and recall with respect to Merriam-Webster for this example are 18.4% and 21.2%, respectively.[13] However, several considerations soften the effect of these low figures. In the first place, 23 of the R-words that are specific to Mirrors clearly could plausibly have been included in the Merriam-Webster entry; i.e., Merriam-Webster does not give the complete picture of the language. Assuming these entries added to Merriam-Webster would have increased the precision to 78.9%. In the second place 7 of the R-words specific to Webster do not occur in the ENPC, which means that their absence does not reflect a shortcoming of the method. Disregarding those words increases recall to 26.9% In the third place, the relationship between the two standard resources Merriam-Webster and Princeton WordNet is not strikingly different from the relationship between Mirrors and them, a fact clearly suggesting that neither of them can be considered as a complete, universal gold standard.

5 CONCLUSION

Semantic relations among the words of a language to some extent vary with the domain of discourse. But the amount of labour required to build domain-specific resources of this kind manually is forbidding. This creates a need for automatized methods of the kind presented here, which in turn underscores the importance of developing high-quality parallel corpora.

REFERENCES

Dyvik, Helge 1998a. A translational basis for semantics. In Stig Johansson and Signe Oksefjell (eds.): *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi 1998, 51-86.

Dyvik, Helge 1998b. Translations as semantic mirrors. In *Proceedings of Workshop W13: Multilinguality in the lexicon II, pp. 24.44,* Brighton, UK. The 13th biennial European Conference on Artificial Intelligence ECAI 98.

Dyvik, Helge 2002. Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers*, 1 April 2004, vol. 49, iss. 1, 311-326(16) Rodopi.

Dyvik, Helge 2005. Translations as a semantic knowledge source. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, Institute of Cybernetics, Tallinn University of Technology, and Institute of the Estonian Language,Tallinn, 27-38.

Ide, Nancy 1999a. Word sense disambiguation using cross-lingual information. In *Proceedings of ACH-ALLC '99 International Humanities Computing Conference*, Charlottesville, Virginia. http://jefferson.village.virginia.edu/ach-allc.99/proceedings

Ide, Nancy 1999b. Parallel translations as sense discriminators. In *SIGLEX99: Standardizing Lexical Resources, ACL99 Workshop*, College Park, Maryland, 52-61.

Ide, Nancy, Tomas Erjavec & Dan Tufis 2002. Sense Discrimination with Parallel Corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, 54-60.

Johansson, Stig, Jarle Ebeling, and Knut Hofland 1996. Coding and aligning the English-Norwegian Parallel Corpus. In K. Aijmer, B. Altenberg, and M. Johansson (eds.) 1996. *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*, 73-85. Lund: Lund University Press, 87-112.

Johansson, Stig, Jarle Ebeling, and Signe Oksefjell 1999/2002. *English-Norwegian Parallel Corpus: Manual.* Department of British and American Studies, University of Oslo.

Lyse, Gunn Inger 2003. *Fra speilmetoden til automatisk ekstrahering av et betydningstagget korpus for WSD-formål.* Masters thesis, University of Bergen.

Pianta, E. and Luisa Bentivogli 2003. Translation as Annotation. In *Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy", Pisa, Italy, September 23-26, 2003:* 40-48.

Priss, Uta and John Old 2005. Conceptual Exploration of Semantic Mirrors. In Ganter, Godin (eds.) *Formal Concept Analysis: Third International Conference, ICFCA 2005*, Springer Verlag.

Thunes, Martha 2003. Evaluating thesaurus entries derived from translational features. Paper presented at *Nodalida 2003*, Reykjavík.

Tufis, Dan and Radu Ion 2003. Word sense clustering based on translation equivalence in parallel texts; a case study in Romanian. In *Proceedings of the International Conference on Speech and Dialog – SPED*, 13-26, Bucharest.

Tufis, Dan, Radu Ion, Nancy Ide 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned

[13] The figures for the full set of 43 adjectives are 18.5% and 13.5%, respectively.

Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, 1312-1318, Geneva.

Resnik, Philip Stuart & David Yarowsky 1997. A perspective on word sense disambiguation methods and their evaluation, position paper presented at the ACL SIGLEX Workshop on *Tagging Text with Lexical Semantics: Why, What, and How?*, held April 4-5, 1997 in Washington, D.C., USA in conjunction with ANLP-97. [11] E. Freeman, S. Hupfer, and K. Arnold. *JavaSpaces: Principles, Patterns and Practice*. Addison-Wesley, USA. (1999)