Automatic Selection and Ranking of Translation Candidates

Antonio Sanfilippo, Ralf Steinberger

SHARP Laboratories of Europe, Ltd.
Oxford Science Park
Oxford OX4 4GA, England
{antonio,ralf}@sharp.co.uk

Abstract. We propose a method for selecting and ranking translation candidates using as, input disambiguated source language expressions with thesaurus-compatible senses. This procedure provides the means for choosing contextually appropriate translations automatically once the sense of the expression in the source language is known. Results can be stored to create a database where bilingual dictionary entries are enriched with information from monolingual thesauri. Bilingual dictionaries enriched with thesaurus information can also be used to generate dictionaries for new language pairs.

1 Introduction

One of the major problems in multilingual NLP applications is to identify which of the various translation candidates for an expression is appropriate in a given context. For example, the English word *bad* translates into German *böse* in the 'immoral' sense, *krank* in the 'unhealthy' sense and *faul* or *verdorben* in the 'rotten' sense, to mention just a few. It is obvious that in the absence of sense information it is not possible to choose the appropriate German translation for *bad* in an expression such as *bad leg* automatically.

Although bilingual dictionaries provide some information about word sense distinctions, the sense classifications in them tend to differ from those found in monolingual dictionaries. For example, the adjective *bad* has 11 senses in the *Concise Oxford Thesaurus* [Kirkpatrick, 1995], whereas it has 8-10 in the *English/German* [Scholze-Stubenrecht & Sykes, 1990], *English/Spanish* [Jarman Galimberti & Russell, 1994] and *English/French Oxford Bilingual Dictionaries* [Corréard & Grundy, 1994], These sense classification mismatches are a major problem for multilingual NLP applications where the input to bilingual dictionary lookup is often the result of grammatical analysis using monolingual dictionaries. If word senses in the monolingual dictionaries do not match the word sense distinctions made in bilingual dictionaries, an informed translation choice cannot be readily made.

There are now known techniques which make it possible to carry out word disambiguation in context using machine readable thesauri (see [Yarowsky, 1995, Resnik, 1995, Sanfilippo, 1997] and references therein). For example, the disambiguation of the adjective *bad* in the context *to cure a bad leg* will provide a specific sense for *bad* (see Fig 1) relative to a machine readable thesaurus of choice (see Fig 2). These techniques can be used to help identify the contextually appropriate translation automatically, provided that sense information in the thesaurus and the bilingual dictionary is compatible. Suppose, for example, we knew that *bad* in the context *to cure a bad leg* corresponds to sense 9 of the thesaurus entry in Fig 2. The correct German translation would then be easily retrieved from bilingual entries such as those in Fig 3 where the integer indicates the relevant thesaurus word sense. The method described in this paper satisfies

these sense compatibility requirements by providing a technique for linking senses across the monolingual thesaurus and bilingual dictionary.

INPUT: < cure_v bad_adj leg_noun >
OUTPUT: < ... bad_adj_[sns:9] ... >

Figure 1. Disambiguation of adjective in context

bad *adj* 1 *bad workmanship|a bad driver* poor, unsatisfactory, inadequate, deficient, imperfect, defective, inferior, substandard, faulty, unacceptable, useless, worthless, inept, ineffectual, duff, ropy. 2 *smoking is a bad habit|its bad for you* harmful, hurtful, damaging, dangerous, injurious, detrimental, destructive, ruinous, deleterious, unhealthy, unwholesome, poisonous. 3 *a bad man|leading a bad life* immoral, wicked, wrong, evil, sinful, corrupt, base, reprobate, depraved, dishonest, dishonourable, crooked. 4 *a bad child* naughty, mischievous, disobedient, unruly, wayward, refractory. 5 *bad weather|having a bad time* disagreeable, unpleasant, unwelcome, uncomfortable, nasty, terrible, dreadful, adverse, grim, gloomy, unfortunate, unfavourable, unlucky, distressing. 6 *a bad time for house-buying* adverse, difficult, unfavourable, unfortunate, unsuitable, inappropriate, inapt. 7 *a bad mistake/accident* serious, severe, grave, disastrous, terrible, critical, acute. 8 *bad eggs|meat going bad* rotten, off, decayed, mouldy, putrid, tainted, spoiled, contaminated, putrescent, putrefacient. 9 *an invalid feeling bad* ill, unwell, sick, poorly, indisposed, ailing, weak, feeble, diseased, under the weather, below par. 10 *feeling bad about their actions* sorry, apologetic, regretful, conscience-stricken, contrite, remorseful, guilty, penitent, rueful, sad, upset. 11 *a bad cheque* worthless, invalid, counterfeit, false, spurious, fraudulent, fake, bogus, phoney.

Figure2. Thesaurus entry for the adjective *bad* (adapted from [Kirkpatrick, 1995]).

bad_adj_[sns:3] ⇔ böse_adj ('immoral' sense)
bad_adj_[sns:9] ⇔ krank_adj ('unhealthy' sense)

Figure 3. Bilingual entries with sense information

The practice of linking word senses across bilingual dictionaries and thesauri is also crucial in the compilation of bilingual thesauri which has become the focus of much research in computational lexicography, see [CRISTAL, 1993] and [EuroWordNet, 1996]. Most existing methods (e.g. [Rigau & Agirre, 1996} and [Knight & Luk, 1994] tend to rely on hierarchically structured thesauri such as WordNet [Miller, 1990] to guide sense linking. One exception to this trend is an approach developed by (Okumura & Hovy, 1994] where overlap of sübcategorization information and words in example sentences is used as the guiding criterion. Our method differs from these approaches in that it only uses a machine readable dictionary of synonyms as the monolingual lexical database (i.e. synonym sets need not be hierarchically structured), and it requires no information about complementation patterns. This reduction in source knowledge requirements is of significant value as (*i*) hierarchically structured thesauri are not easily available and very

costly to produce, and (*ii*) detailed knowledge about complementation patterns is ordinarily not available in bilingual dictionaries. Of course, it is also possible to combine different approaches so as to reduce the chance for errors and/or the occurrence of linking failures.

2 Linking senses across bilingual dictionaries and thesauri

The method described provides the means for selecting and ranking translation candidates automatically. It uses machine readable bilingual dictionaries and monolingual thesauri as knowledge sources. The input data are (lemmatized) lexical units which have been disambiguated through the assignment of word senses from the relevant monolingual thesaurus used as knowledge source. In the present paper, we will consider an example where the Source Language (SL) is English, the Target Language (TL) is German, and the knowledge sources used by the system are the English thesaurus and English/German bilingual dictionary fragments in Fig 4 and Fig 5.

bad *adj* **3** *a bad man|leading a bad life,* immoral, wicked, wrong, evil, sinful, corrupt, base, depraved, dishonest, dishonourable, crooked. **9** *an invalid feeling bad* ill, unwell, sick, poorly, indisposed, ailing, weak, feeble, diseased.

Figure 4. Thesaurus fragment

Given as input the disambiguated English word **bad_adj_[sns:9]**, the system will give as output the set of translation candidates **{krank_adj_[cm:.5]}, {schwach_adj_[cm:.25]}, {schlecht_adj_[cm: .25]}** where **cm:** followed by an integer provides a numerical reading of the confidence measure.

The merger of an equivalence from a bilingual dictionary with its corresponding entry in a monolingual thesaurus is carried out in 3 main steps.

2.1 Step 1: Retrieving synonyms of the SL word

Given as input a sense-disambiguated expression — where disambiguation is carried out through the assignment of word senses from the monolingual thesaurus used in the system (e.g. Fig 4). — the first simple step involves retrieving the synonyms for the relative sense of the expression from the relevant thesaurus (Fig 4), e.g.

INPUT: bad_adj_[sns:9]

OUTPUT: {ill_adj, unwell_adj, sick_adj, poorly_adj, indisposed_adj, ailing_adj, weak_adj, feeble_adj, diseased_adj}

2.2 Step 2: Translating synonyms in the target language

In the next step, the TL translations for the synonyms retrieved in step 1 and for the input expression itself should be retrieved using the bilingual dictionary (Fig 4). The result is a large set of words including all potential translation candidates. Note that several translation candidates occur more than once. These are the words which are the most appropriate translations

bad_adj ⇔ arg_adj
bad_adj ⇔ böse_adj
bad_adj ⇔ faul_adj
bad_adj ⇔ grob_adj
bad_adj ⇔ krank_adj
bad_adj ⇔ schlecht_adj
bad_adj ⇔ schlimm_adj
bad_adj ⇔ schwer_adj
bad_adj ⇔ stark_adj
bad_adj ⇔ unartig_adj
bad_adj ⇔ unrein_adj
bad_adj ⇔ übel_adj
immoral_adj ⇔ sittenlos_adj
immoral_adj ⇔ unmoralisch_adj
wicked_adj ⇔ boshaft_adj
wicked_adj ⇔ böse_adj
wicked_adj ⇔ frech_adj
wicked_adj ⇔ frevelhaft_adj
wicked_adj ⇔ schlecht_adj
wrong_adj ⇔ falsch_adj
wrong_adj ⇔ fälschlich_adj
wrong_adj ⇔ unrecht_adj
wrong_adj ⇔ verkehrt_adj
evil_adj ⇔ böse_adj
sinful_adj ⇔ sündhaft_adj
corrupt_adj ⇔ korrupt_adj
corrupt_adj ⇔ verdorben_adj
base_adj ⇔ gemein_adj
base_adj ⇔ niederträchtig_adj
base_adj ⇔ niedrig_adj
base_adj ⇔ unedel_adj
depraved_adj ⇔ lasterhaft_adj
depraved_adj ⇔ verdorben_adj
depraved_adj ⇔ verkommen_adj
depraved_adj ⇔ verwahrlost_adj
dishonest_adj ⇔ unehrlich_adj
dishonest_adj ⇔ unlauter_adj
dishonest_adj ⇔ unsauber_adj
dishonourable_adj ⇔ ehrlos_adj
dishonourable_adj ⇔ unehrenhaft_adj
crooked_adj ⇔ krumm_adj
crooked_adj ⇔ schief_adj
crooked_adj ⇔ unehrlich_adj
crooked_adj ⇔ verkrümmt_adj
ill_adj ⇔ schlecht_adj
ill_adj ⇔ krank_adj
unwell_adj ⇔ unwohl_adj
sick_adj ⇔ makaber_adj
sick_adj ⇔ krank_adj
poorly_adj ⇔ elend_adj
indisposed_adj ⇔ indisponiert_adj
indisposed_adj ⇔ unpäßlich_adj
ailing_adj ⇔ kränkelnd_adj
weak_adj ⇔ dünn_adj
weak_adj ⇔ kraftlos_adj
weak_adj ⇔ matt_adj
weak_adj ⇔ schwach_adj
weak_adj ⇔ weichlich_adj
feeble_adj ⇔ schwach_adj
diseased_adj ⇔ krank_adj

Figure 5. Bilingual Dictionary fragment

for the input word in the given sense. Out of this set of words, only the translations which exhibit repeated occurrences should be selected:

INPUT1: {ill_adj, unwell_adj, sick_adj, poorly_adj, indisposed_adj, ailing_adj, weak_adj, feeble_adj, diseased_adj}
INPUT2: bad_adj

INTERMEDIATE RESULT: {arg_adj, böse_adj, faul_adj, grob_adj , krank_adj, schlecht_adj , schlimm_adj , schwer_adj , stark_adj , unartig_adj , unrein_adj , übel_adj, schlecht_adj, krank_adj , unwohl_adj , makaber_adj , krank_adj , elend_adj, indisponiert_adj , unpäßlich_adj , kränkelnd_adj, dünn_adj , kraftlos_adj , matt_adj, schwach_adj , weichlich_adj , schwach_adj , krank_adj}

OUTPUT: {krank_adj, schlecht_adj, schlecht_adj, krank_adj, krank_adj, schwach_adj, schwach_adj, krank_adj}

2.3 Step 3: Scoring translations

The translations in the resulting set of step 2 can now be scored according to their occurrence rate. The idea is that the more frequent the translation is the more appropriate it is. The frequency of occurrence can thus be seen as a confidence measure expressing the appropriateness of this translation for the SL word in its given sense.

INPUT: {schlecht_adj, krank_adj, krank_adj, schwach_adj, schwach_adj, krank_adj, krank_adj, schlecht_adj}

OUTPUT: {krank_adj_[cm:4], schwach_adj_[cm:2] , schlecht_adj_[cm:2]}

It is useful to express the confidence measure (cm) in normalised form. Confidence measure scores can be normalized by dividing each score by the sum of all scores using the formula:

$$norm(cm_i) = \frac{cm_i}{\sum_1^n cm}$$

For example, **krank** is assigned a confidence measure of 0.5 as it has 4 occurrences and the sum of all translation occurrences is 8 (i.e. 4 for **krank**, 2 for **schwach** and 2 for **schlecht**).

OUTPUT: {krank_adj_[cm:0.5], schwach_adj_[cm:0.25], schlecht_adj.[cm:0.25]}

2.4 Creation of new equivalences

An interesting fact is that this process may provide new translations for the SL word. In the example given above, the German word schwach was not given in any of the bilingual equivalences for the word bad in Fig 5. Although the new translations found in this way are very likely to be appropriate translations of the given SL word, a user may not want to include new translations. In this case it is easy to exclude new translations by comparing the result with the initial set of bilingual equivalences.

2.5 Coping with failures

In certain cases, the method presented in 2.1 to 2.3 does not yield any results, namely when

1. no synonyms were found in step 1,
2. there are no translations for the synonyms retrieved,
3. there are no repeated occurrences among translated synonyms.

In the first case the translations of the input expression are returned as output; the result obtained is equivalent to a regular lookup in the bilingual dictionary.

In the remaining two cases, a further attempt at selecting a translation candidate for a word sense *WS* can be made by repeating steps 2 and 3 taking as input the word set containing the synonyms of each synonym *Syn* of *WS* for each sense of *Syn*. If still no result is obtained, the process can be repeated taking as input the synonyms of the synonyms of the synonyms of *WS* and so on. For each iteration, the distance between *WS* and the synonyms whose translations are used to select and rank translation candidates for *WS* is recorded. A greater distance will introduce more noise in the selection process and will thus be regarded as lowering the confidence measure.

3 Storing results

The results of the third step (§2.3) can be stored together with the input expression in the form of a bilingual equivalence, e.g.

INPUT1: bad_adj_[sns:9]
INPUT2: {krank_adj_[cm:0.5], schwach_adj_[cm:0.25], schlecht_adj_[cm:0.25]}

OUTPUT: bad_adj_[sns:9] ⇔ {krank_adj_[cm:0.5],
schwach_adj_[cm:0.25],
schlecht_adj_[cm:0.25]}

Repeating the procedure in §2.1-§2.3 for all sense entries of all words in the thesaurus results in bilingual dictionaries enriched with thesaurus information which can be used in a variety of multilingual NLP applications. However, it may not always be possible to carry out this pre-processing, but the procedure may have to be carried out at run-time. This could be the case, for instance, when users are allowed to plug their own bilingual dictionaries into a multilingual NLP system. When the selection and ranking of the translation candidates is done at run-time, the stored results of the first uses can be consulted in subsequent uses of the system so as to avoid repeating the same selection and ranking processes. In this way it is possible to incrementally build a bilingual lexical database enriched with thesaurus information.

4 Linking Bilingual Dictionaries

Given three languages A, B and C, it is often the case that bilingual dictionaries are available for language pairs A/C and A/B and missing for B/C. When such a situation arises, it might be possible to form equivalences for the missing language pair linking the B and C sides of equivalences in the A/B and A/C bilingual dictionaries which have the same A expression. Suppose, for example, that we have the translation of English *liver* and *knee* into German and Italian; all we need to do in order to create German/Italian equivalences for the two words is to relate Italian and German expressions which have the same English translation, e.g.

INPUT: {E/I: liver_noun ⇔ fegato_noun,
E/I: knee_noun ⇔ ginocchio_noun,
E/G: liver_noun ⇔ Leber_noun,
E/G: knee_noun ⇔ Knie_noun}

OUTPUT: (G/I: Leber_noun ⇔ fegato _noun,
G/I: Knie_noun ⇔ ginocchio_noun}

In this example, the situation is simplified by the fact that the English words are not ambiguous: there is only one translation for *liver* and *knee* in both German and Italian. If the English words were to have more than one translation in German and/or Italian, it would be impossible to determine the appropriate G/I equivalences solely with reference to the English side of the equivalences. For example, given the E/I and E/G equivalences below, this simplistic procedure would produce the G/I equivalence **böse_adj ⇔ malato_adj**, which is clearly wrong as it equates the 'immoral' and 'unhealthy' senses of *bad.*

E/I: bad_adj ⇔ malato_adj
E/I: bad_adj ⇔ cattivo_adj
E/G: bad_adj ⇔ krank_adj
E/G; bad_adj ⇔ böse_adj

The method described provides a solution to this problem as it makes it possible to distinguish homonyms by assignment of thesaurus senses. For example, we could apply the method described in §2 to assign thesaurus senses to the E/G and E/I bilingual entries as follows:

INPUT: {bad_adj_[sns:3],
bad_adj_[sns:9]}

OUTPUT1: {E/G: bad_adj_[sns:9] ⇔ krank_adj,
E/G: bad_adj_[sns:3] ⇔ böse_adj>
OUTPUT2: {E/I: bad_adj_[sns:9] ⇔ malato_adj,
E/I: bad_adj_[sns:3] ⇔ cattivo_adj)

Such an assignment will make it possible to determine the appropriate G/I equivalences solely with reference to the English side of bilingual entries since each G/I translation has a unique association with one of the senses of *bad,* e,g.

INPUT: {E/I: bad_adj_[sns:9] ⇔ malato_adj,
E/I: bad_adj_[sns:3] ⇔ cattivo_adj,
E/G: bad_adj_[sns:9] ⇔ krank_adj,
E/G: bad_adj_[sns:3] ⇔ böse_adj>

OUTPUT: {G/I: krank_adj ⇔ malato_adj,
G/I: böse_adj_[sns:3] ⇔ cattivo_adj>

Of course, bilingual dictionaries do often make sense distinctions, e.g. by grouping translations which refer to the same sense of the SL expression. However, the sense distinctions made by a bilingual dictionary for a given language are not likely to coincide with those made by another bilingual dictionary for the language. Therefore, the use of a monolingual thesaurus as a general reference point for sense assignment is necessary to identify common sense readings for the same language across bilingual dictionaries.

5 Evaluation

We evaluated the method described with reference to 21593 noun entries which occurred in both the English/German bilingual dictionary [Scholze-Stubenrecht & Sykes, 1990] and the English thesaurus [Kirkpatrick, 1995].[1] On average, the nouns have 1,8 readings.

Table 1 shows recall and precision relative to the assignment of thesaurus readings to bilingual entries. The first column reports results obtained with the synonym set of the input word, as described in §2.1-§2.3. The second column reports results obtained using the word set containing the synonyms of the synonyms of the input word, as described in §2.5. Recall is considerably higher when using expanded synonym sets (second column), while precision is somewhat lower. The degradation of precision is due to the introduction of false near synonyms in expanded synonym sets which follows from retrieving the synonyms of all senses of the synonym of the input word (see §2.5). Notice that recall could not be 100 % due to the presence of sense gaps across the bilingual dictionary and thesaurus. Many such instances were correctly identified.

Table 2 reports the recall and precision in ranking translations of a given noun sense. Recall provides a measure of how often thesaurus senses were ranked with reference to senses in the bilingual dictionary; precision specifies how often such rankings were correctly made. As in the previous case, the use of extended synonym sets triggers an increase in recall and a decrease in precision. In both cases, this effect could be greatly reduced by using a thesaurus with sense-disambiguated synonyms,

	Original Synonym Set	Expanded Synonym Set	Number of senses
Recall	48%	97%	21593
Precision	100%	86%	104

Table 1. Linking senses across the bilingual dictionary and thesaurus

	Original Synonym Set	Expanded Synonym Set	Number of senses
Recall	44%	92%	751
Precision	100%	87%	69

Table 2. Ranking translation candidates

[1] One problem we found in relating bilingual dictionaries and thesauri is the presence of lexical gap«, i.e. a word is present in the bilingual dictionary and missing in the thesaurus. This problem may be addressed by using combined lexical sources,

6 Conclusion

Most current approaches to lexical disambiguation make use of thesauri as knowledge sources to carry out word sense discrimination. Consequently, a reliable way of linking bilingual dictionaries to thesauri must be available if bilingual dictionary lookup is to be informed by word sense disambiguation. In this paper, we have proposed to achieve this goal by intersecting translations for the thesaurus synonyms of a word's sense with translations for each sense of the word in a bilingual dictionary. This method provides good results with ordinary lexical database sources, and can be further improved by using a thesaurus with sense-disambiguated synonyms.

References

[CRISTAL, 1993] CRISTAL. 1993, Conceptual Retrieval of Information using Semantic dicTionary in the three Languages, LRE-2 62059, `http://www2.echo.lu/langeng/en/lre2/cristal.htBl`.

[Corrérd & Grundy, 1994] Corrérd, M. & V. Grundy. 1994, *The Oxford Hachette French Dictionary: French-English, English-French,* Oxford University Press, New York.

[EuroWordNet, 1996] EuroWordNet. 1996, *Building a Multilingual Wordnet Database with Semantic Relations between Words,* LE-2 4003, `http://www2.echo.lu/langeng/en/le2/eurowordnet`.

[Jarman Galimberti & Russell, 1994] Jarman Galimberti, B. & R. Russell. 1994, *The Oxford Spanish Dictionary: Spanish-English, English-Spanish,* Oxford University Press, New York.

[Kirkpatrick, 1995] Kirkpatrick, B. 1995, *The Concise Oxford Thesaurus,* Oxford University Press, Oxford.

[Knight & Luk, 1994] Knight, Kevin and Steve K. Luk. 1994, Building a Large-Scale Knowledge Base for Machine Translation, In *Proceedings of AAAI-94,* Stanford.

[Miller, 1990] Miller G. 1990, Five Papers on WordNet, *International Journal of Lexicography,* 3(4) (special issue).

[Okumura & Hovy, 1994] Okumura, Akitoshi and Eduard Hovy. 1994, Lexicon-to-Ontology Concept Association Using a Bilingual Dictionary, In *Proceedings of AMTA-94,* Columbia.

(Resnik, 1995] Resnik, P. 1995, Disambiguating noun groupings with respect to WordNet Senses, In *Proceedings of the Third Workshop on Very Large Corpora.* Cambridge, Mass., pp. 54-68.

[Rigau & Agirre, 1996] Rigau, German and Eneko Agirre. 1996, Linking Bilingual Dictionaries to WordNet, In *Proceedings of Eurolex-96,* Göteborg.

[Sanfilippo, 1997] Sanfilippo, Antonio. 1997, *Using Semantic Similarity to Acquire Cooccurrence Restrictions from Corpora,* To appear in *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications,* Madrid, 1997.

[Scholze-Stubenrecht & Sykes, 1990] Scholze-Stubenrecht, W. & J. B. Sykes. 1990, *The Oxford Duden German Dictionary: German-English, English-German,* Oxford University Press, New York,

[Yarowsky, 1995] Yarowsky D. 1995, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of ACL-95,* pp. 189-196.