# Exploiting Aggregate Properties of Bilingual Dictionaries For Distinguishing Senses of English Words and Inducing English Sense Clusters

**Charles SCHAFER and David YAROWSKY**
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD, 21218, USA
{cschafer,yarowsky}@cs.jhu.edu

## Abstract

We propose a novel method for inducing monolingual semantic hierarchies and sense clusters from numerous foreign-language-to-English bilingual dictionaries. The method exploits patterns of non-transitivity in translations across multiple languages. No complex or hierarchical structure is assumed or used in the input dictionaries: each is initially parsed into the "lowest common denominator" form, which is to say, a list of pairs of the form (foreign word, English word). We then propose a monolingual synonymy measure derived from this aggregate resource, which is used to derive multilingually-motivated sense hierarchies for monolingual English words, with potential applications in word sense classification, lexicography and statistical machine translation.

## 1 Introduction

In this work we consider a learning resource comprising over 80 foreign-language-to-English bilingual dictionaries, collected by downloading electronic dictionaries from the Internet and also scanning and running optical character recognition (OCR) software on paper dictionaries. Such a diverse parallel lexical data set has not, to our knowledge, previously been assembled and examined in its aggregate form as a lexical semantics training resource. We show that this aggregate data set admits of some surprising applications, including discovery of synonymy relationships between words and automatic induction of high-quality hierarchical word sense clusterings for English.

We perform and describe several experiments deriving synonyms and sense groupings from the aggregate bilingual dictionary, and subsequently suggest some possible applications for the results.

Finally, we propose that sense taxonomies of the kind introduced here, being of different provenance from those produced explicitly by lexicographers or using unsupervised corpus-driven methods, have significant value because they add diversity to the set of available resources.

## 2 Resources

First we collected, from Internet sources and via scanning and running OCR on print dictionaries, 82 dictionaries between English and a total of 44 distinct foreign languages from a variety of language families.

Over 213K distinct English word types were present in a total of 5.5M bilingual dictionary entries, for an average of 26 and a median of 3 foreign entries per English word. Roughly 15K English words had at least 100 foreign entries; over 64K had at least 10 entries.

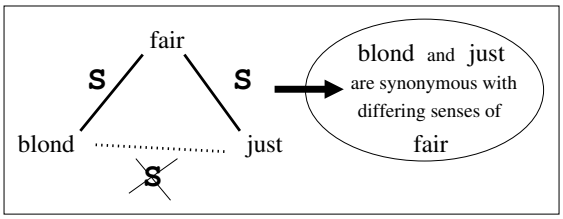


Figure 1: Detecting asynonymy via unbalanced synonymy relationships among 3 words. The derived synonymy relation S holds between *fair* and *blond*, and between *fair* and *just*. S does not hold between *blond* and *fair*. We can infer that *fair* has at least 2 senses and, further, we can represent them by *blond* and *just*.

| English | French | Spanish | German |
|---|---|---|---|
| *fair* | blond, juste | blondo, licito, recto | blond, gerecht |
| blond | blond | blondo | blond |
| just | juste | licito; recto | gerecht |

Figure 2: This excerpt from the data set illustrates the kind of support the aggregate bilingual dictionary provides for partitioning the meanings of *fair* into distinct senses: *blond* and *just*.

No complex or hierarchical structure was assumed or used in our input dictionaries. Each was initially parsed into the "lowest common denominator" form. This consisted of a list of pairs of the form (foreign word, English word). Because bilingual dictionary structure varies widely, and even the availability and compatibility of part-of-speech tags for entries is uncertain, we made the decision to compile the aggregate resource only with data that could be extracted from every individual dictionary into a universally compatible format. The unique pairs extracted from each dictionary were then converted to 4-tuples of the form:

*<foreign language, dictionary name, foreign word, English word>*

before being inserted into the final, combined dictionary data set.

## 3 A Synonymy Relation

We began by using the above-described data set to obtain a synonymy relation between English words.

In general, in a paper bilingual dictionary, each for-

eign word can be associated with a list of English words which are possible translations; in our reduced format each entry lists a single foreign word and single possible English translation, though taking a union of all English translations for a particular foreign word recreates this list.

We use the notion of *coentry* to build the synonymy relation between English words. The per-entry coentry count $C_{per-entry}(e_1,e_2)$ for two English words $e_1$ and $e_2$ is simply the number of times $e_1$ and $e_2$ both appear as the translation of the same foreign word (over all foreign words, dictionaries and languages). The per-dictionary coentry count $C_{per-dict}(e_1,e_2)$, ignores the number of individual coentries within a particular dictionary and merely counts as 1 any number of coentries inside a particular dictionary. Finally, per-language coentry count $C_{per-lang}(e_1,e_2)$ counts as 1 any number of coentries for $e_1$ and $e_2$ for a particular language. Thus, for the following snippet from the database:

| Eng. Wd. | Foreign Wd. | Foreign Language | Dict. ID |
|---|---|---|---|
| hit | schlagen | GERMAN | ger.dict1 |
| pound | schlagen | GERMAN | ger.dict1 |
| hit | schlag | GERMAN | ger.dict1 |
| pound | schlag | GERMAN | ger.dict1 |
| hit | schlag | GERMAN | ger.dict2 |
| pound | schlag | GERMAN | ger.dict2 |
| hit | battere | ITAL | ital.dict1 |
| pound | battere | ITAL | ital.dict1 |

$C_{per-entry}$(*hit,pound*) = 4, while $C_{per-dict}$(*hit,pound*) = 3, since the two individual coentries in *ger.dict1* are only counted once. $C_{per-lang}$(*hit,pound*) = 2; *hit* and *pound* are coentries in the Italian and German languages. We found the more conservative per-dictionary and per-language counts to be a useful device, given that some dictionary creators appear sometimes to copy and paste identical synonym sets in a fairly indiscriminate fashion, spuriously inflating the $C_{per-entry}(e_1,e_2)$ counts.

Our algorithm for identifying synonyms was simple: we sorted all pairs of English words by decreasing $C_{per-dict}(e_1,e_2)$ and, after inspection of the resulting list, cut it off at a per-dictionary and per-language count threshold[1] yielding qualitatively strong results. For all word pairs $e_1$,$e_2$ above threshold, we say the symmetric synonymy relation S($e_1$,$e_2$) holds. The following tables provide a clarifying example showing how synonymy can be inferred from multiple bilingual dictionaries in a way which is impossible with a single such dictionary (because of idiosyncratic foreign language polysemy).

| Lang. | Dict. ID | Foreign Wd | English Translations |
|---|---|---|---|
| GERMAN | ger.dict1 | absetzen | **deposit** drop deduct **sell** |
| GERMAN | ger.dict1 | ablagerung | **deposit sediment** settlement |

The table above displays entries from one German-English dictionary. How can we tell that "sediment" is a better synonym for "deposit" than "sell"? We can build and examine the coentry counts $C_{per-lang}$(*deposit,sediment*) and $C_{per-lang}$(*deposit,sell*) using dictionaries from many languages, as illustrated below:

| | | | |
|---|---|---|---|
| FRENCH | fre.dict1 | dépôt | arsenal **deposit** depository depot entrusting filing sludge store trust submission repository scale **sediment** |
| TURKISH | tk.dict1 | tortu | **sediment deposit** faeces remainder dregs crust |
| CZECH | cz.dict1 | sedlina | clot **deposit sediment** warp |

Polysemy which is specific to German – "deposit" and "sell" senses coexisting in a particular word form "absetzen" – will result in total coentry counts $C_{per-lang}$(*deposit,sell*), over all languages and dictionaries, which are low. In fact, "deposit" and "sell" are coentries under only 2 out of 44 languages in our database (German and Swedish, which are closely related). On the other hand, near-synonymous English translations of a particular sense across a variety of languages will result in high coentry counts, as is the case with $C_{per-lang}$(*deposit,sediment*). As illustrated in the tables, German, French, Czech and Turkish all support the synonymy hypothesis for this pair of English words.

| "deposit" Coentries | Per Entry | Per Dict. | Per Lang. |
|---|---|---|---|
| sell | 4 | 4 | 2 |
| sediment | 68 | 40 | 18 |

The above table, listing the various coentry counts for "deposit", demonstrates the empirical motivation in the aggregate dictionary for the synonymy relationship between *deposit* and *sediment*, while the aggregate evidence of synonymy between *deposit* and *sell* is weak, limited to 2 languages, and is most likely the result of a word polysemy restricted to a few Germanic languages.

## 4 Different Senses: Asymmetries of Synonymy Relations

After constructing the empirically derived synonymy relation S described in the previous section, we observed that one can draw conclusions from the topology of the graph of S relationships (edges) among words (vertices).

Specifically, consider the case of three words $e_1$,$e_2$, $e_3$ for which S($e_1$,$e_2$) and S($e_1$,$e_3$) hold, but S($e_2$,$e_3$) does not. Figure 1 illustrates this situation with an example from data ($e_1$ = "fair"), and more examples are listed in Table 1. As Figure 1 suggests and inspection of the random extracts presented in Table 1 will confirm, this topology can be interpreted as indicating that $e_2$ and $e_3$ exemplify differing senses of $e_1$.

We decided to investigate and apply it with more generality. This will be discussed in the next section.

## 5 Inducing Sense Taxonomies: Clustering with Synonym Similarity

With the goal of using the aggregate bilingual dictionary to induce interesting and useful sense distinctions of English words, we investigated the following strategy.

[1]The threshold was 10 and 5 respectively for per-dictionary and per-language coentry counts.

| $syn_1$(W) | W | $syn_2$(W) |
|---|---|---|
| quiet | **still** | yet |
| desire | **want** | lack |
| delicate | **tender** | offer |
| conceal | **hide** | skin |
| nice | **kind** | sort |
| assault | **charge** | load |
| filter | **strain** | stretch |
| flow | **run** | manage |
| cloth | **fabric** | structure |
| blond | **fair** | just |
| foundation | **base** | ignoble |
| deny | **decline** | fall |
| hurl | **cast** | mould |
| bright | **clear** | open |
| harm | **wrong** | incorrect |
| crackle | **crack** | fissure |
| impeach | **charge** | load |
| enthusiastic | **keen** | sharp |
| coarse | **rough** | difficult |
| fling | **cast** | form |
| firm | **fast** | speedy |
| fashion | **mold** | mildew |
| incline | **lean** | meagre |
| arouse | **raise** | increase |
| digit | **figure** | shape |
| dye | **paint** | picture |
| spot | **stain** | tincture |
| shape | **cast** | toss |
| claim | **call** | shout |
| earth | **ground** | groundwork |
| associate | **fellow** | guy |
| arrest | **stop** | plug |

Table 1: A representative sampling of high-confidence sense distinctions derived via unbalanced synonymy relationships among three words, W and two of its synonyms $syn_1$(W) & $syn_2$(W), such that $C_{per-dict}$(W,$syn_1$(W)) and $C_{per-dict}$(W,$syn_2$(W)) are high, whereas $C_{per-dict}$($syn_1$(W),$syn_2$(W)) is low (0). Extracted from a list sorted by descending $C_{per-dict}$(W,$syn_1$(W)) $*$ $C_{per-dict}$(W,$syn_2$(W)) / $C_{per-dict}$($syn_1$(W),$syn_2$(W)) (counts were smoothed to prevent division by zero).

For each target word $W_t$ in English having a sufficiently high dictionary occurrence count to allow interesting results[2], a list of likely synonym words $W_s$ was induced by the method described in Section 3[3]. Additionally, we generated a list of all words $W_c$ having nonzero $C_{per-dict}$($W_t$,$W_c$).

The synonym words $W_s$ – the sense exemplars for target words $W_t$ – were clustered based on vectors of coentry counts $C_{per-dict}$($W_s$,$W_c$). This restriction on vector dimension to only words that have nonzero coentries with the target word helps to exclude distractions such as coentries of $W_s$ corresponding to a sense which doesn't overlap with $W_t$. The example given in the following table shows an excerpt of the vectors for synonyms of **strike**. The **hit** synonym overlaps **strike** in the *beat/bang/knock* sense. Restricting the vector dimension as described will help prevent noise from **hit**'s common *chart-topper/recording/hit_single* sense. The following table also illustrates the clarity with which major sense distinctions are reflected in the aggregate dictionary. The induced clustering for **strike** (tree as well as flat cluster boundaries) is presented in Figure 4.

| | attack | bang | hit | knock | walkout | find |
|---|---|---|---|---|---|---|
| attack | - | 4 | 18 | 7 | 0 | 0 |
| bang | - | 38 | 43 | 2 | 0 | 0 |
| hit | | | - | 44 | 2 | 29 |
| knock | | | | - | 2 | 0 |
| walkout | | | | | - | 0 |
| find | | | | | | - |

We used the CLUTO clustering toolkit (Karypis, 2002) to induce a hierarchical agglomerative clustering on the vectors for $W_s$. Example results for **vital** and **strike** are in Figures 3 and 4 respectively[4]. Figure 4 also presents flat clusters automatically derived from the tree, as well as a listing of some foreign words associated with particular clusters.

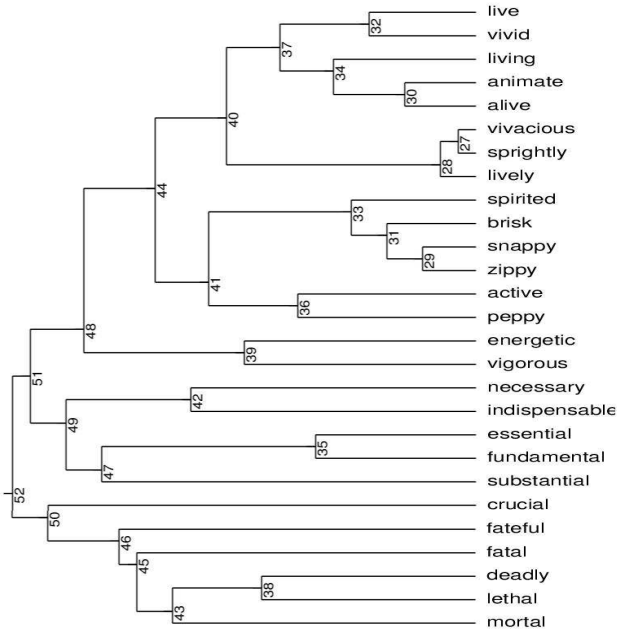


Figure 3: Induced sense hierarchy for the word "**vital**"

## 6 Related Work

There is a distinguished history of research extracting lexical semantic relationships from bilingual dictionaries (Copestake et al., 1995; Chen and Chang, 1998). There is also a long-standing goal of mapping translations and senses in multiple languages in a linked ontology structure (Resnik and Yarowsky, 1997; Risk, 1989; Vossen, 1998). The recent work of Ploux and Ji (2003) has some similarities to the techniques presented here in that it considers topological properties of the graph of synonymy relationships between words. The current paper can be distinguished on a number of dimensions, including our much greater range of participating languages, and the fundamental algorithmic linkage between multilingual translation distributions and monolingual synonymy clusters.

[2]For our experiments, English words occurring in at least 15 distinct source dictionaries were considered.

[3]Again, the threshold for synonyms was 10 and 5 respectively for per-dictionary and per-language coentry counts.

[4]In both "vital" and "strike" examples, the rendered hierarchical clusterings were pruned (automatically) in order to fit in this paper.

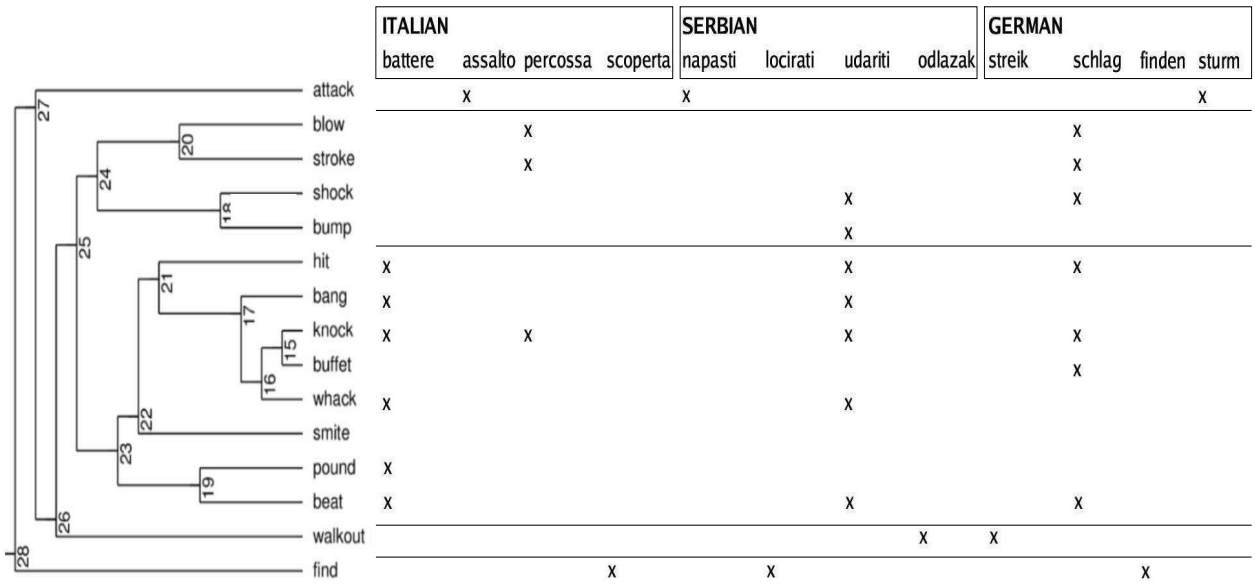


Figure 4: Induced sense hierarchy for the word "**strike**" and some translations of individual "**strike**" synonyms. Flat clusters automatically derived from the tree are denoted by the horizontal lines.

## 7 Analysis and Conclusions

This is the first presentation of a novel method for the induction of word sense inventories, which makes use of aggregate information from a large collection of bilingual dictionaries.

One possible application of the induced sense inventories presented here is as an aid to manual construction of monolingual dictionaries or thesauri, motivated by translation distinctions across numerous world languages. While the desired granularity of sense distinction will vary according to the requirements of taste and differing applications, treating our output as a proposal to be assessed and manually modified would be a valuable labor-saving tool for lexicographers.

Another application of this work is a supplemental resource for statistical machine translation (SMT). It is possible, as shown graphically in Figure 4, to recover the foreign words associated with a *cluster* (not just a single word). Given that the clusters provide a more complete coverage of English word types for a given sense than the English side of a particular bilingual dictionary, clusters could be used to unify bitext co-occurrence counts of foreign words with English *senses* in a way that typical bilingual dictionaries cannot. Unifying counts in this way would be a useful way of reducing data sparsity in SMT training.

Finally, evaluation of induced sense taxonomies is always problematic. First of all, there is no agreed "correct" way to classify the possible senses of a particular word. To some degree this is because human experts disagree on particular judgments of classification, though a larger issue, as pointed out in Resnik and Yarowsky 1997, is that what constitutes an appropriate set of sense distinctions for a word is, emphatically, a function of the task at hand. The sense-distinction requirements of English-to-French machine translation differ from those of English-to-Arabic machine translation (due to differing degrees of parallel polysemy across the language pairs), and both differ from those of English dictionary construction.

We believe that the translingually-motivated word-sense taxonomies developed here will prove useful for the a variety of tasks including those mentioned above. The fact that they are derived from a novel resource, not constructed explicitly by humans or derived in fully unsupervised fashion from text corpora, makes them worthy of study and incorporation in future lexicographic, machine translation, and word sense disambiguation efforts.

## References

J. Chen and J. Chang. 1998. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. *Computational Linguistic*, 29(2):61-95.

A. Copestake, E. Briscoe, P. Vossen, A. Ageno, I. Castellan, F. Ribas, G. Rigau, H. Rodriguez and A. Samiotou. 1995. Acquisition of Lexical Translation Relations from MRDs. *Machine Translation: Special Issue on the Lexicon*, 9(3):33-69.

G. Karypis. 2002. CLUTO: A Clustering Toolkit. *Tech Report 02-017, Dept. of Computer Science, University of Minnesota*. Available at http://www.cs.umn.edu˜cluto

S. Ploux and H. Ji. 2003. A Model for Matching Semantic Maps Between Languages (French/English, English/French). *Computational Linguistics*, 29(2):155-178.

P. Resnik and D. Yarowsky. 1997. A Perspective on Word Sense Disambiguation Methods and Their Evaluation. In *Proceedings of SIGLEX-1997*, pp. 79-86.

O. Risk. 1989. Sense Disambiguation of Word Translations in Bilingual Dictionaries: Trying to Solve The Mapping Problem Automatically. *RC 14666*, IBM T.J. Watson Research Center. Yorktown Heights.

P. Vossen (ed.). 1998. *EUROWORDNET: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers. Dordrecht, The Netherlands.