Retrieving Lexical Semantics from Multilingual Corpora
Ahmad R. Shahid and Dimitar Kazakov

Abstract—This paper presents a technique to build a lexical resource used for annotation of parallel corpora where the tags can be seen as multilingual ‘synsets’. The approach can be extended to add relationships between these synsets that are akin to WordNet relationships of synonymy and hypernymy. The paper also discusses how the success of this approach can be measured. The reported results are for English, German, French, and Greek using the Europarl parallel corpus.

Index Terms—Multilingual corpora, lexical realtions.

I. INTRODUCTION
The aim of this work is to build a WordNet-like resource which can be used for Word Sense Disambiguation (WSD) and other such tasks where semantics of words and phrases is the main objective. The multilingual aspect of the approach helps in reducing the ambiguity inherent in any words/phrases in the pivotal language, which is English in the case shown here.

In order to create such a resource we used proceedings from the European Parliament (Europarl). Four languages were selected with English as the pivotal language in addition to German, French and Greek.

The paragraph-aligned bilingual corpora were fed into a word-alignment tool, GIZA++, to obtain the pair-wise alignments of each language with English. These pair-wise aligned words were later merged into phrases where one word in one language was aligned with more than one word in the other language. Using English as the pivotal language, there were combined into 4-tuples, effectively resulting in a database of multilingual synsets. The synsets were then used to sense disambiguate the individual words and phrases in the original corpora from which they originated. Each of the synsets were latter Part of Speech (POS)-Tagged using the Brill Tagger. The POS tags can help in further removing any ambiguity. Edit distance between any two synsets was also computed in order to use that information for merging any two synsets that are deemed sufficiently close.

II. RELATED WORK
WSD has attracted the attention of the research community for long. It is a tricky issue and needs resources that define the semantic relationships between words. In the last twenty five years various research activities have been undertaken to build large repositories that combined the description of semantic concepts with their relationships. Two efforts worth mentioning here are the Cycorp Cyc project [1] and the lexical semantic database WordNet [2]. Both approaches use a number of predicates to define relationships between concepts, such as “concept A is an instance of concept B” or “concept A is a specific case of concept B.” WordNet also defined the notion of synsets, which defines a semantic concept through all relevant synonyms, e.g., \{mercury, quicksilver, Hg\}.

The original version of the WordNet covered only the English language but the effort has been replicated for other languages as well [3]. Yet all these efforts have been handcrafted, rather than automatically generated and are monolingual in nature. Even though they are highly comprehensive, they require a major, sustained effort to maintain and update.

The work [4] used word alignment in an unsupervised manner to create pseudo-translations which were used for sense tagging of the parallel corpora. They used WordNet as the sense inventory of English. Firstly they aligned each French word with one or more words in English in each sentence. Then to create synsets they looked at the alignment of each French word with all corresponding translations in English in the whole corpus. In order to narrow down the number of combinations they used WordNet to identify nominal compounds, such as honey_bee and queen_bee. WordNet was also used to manually assign sense tags to words in the subset of the corpus used for evaluation. They found the performance of their approach comparable with other unsupervised approaches.

Interest in the use of parallel corpora for unsupervised WSD has grown recently [5], [6]. In both cases, the use of multilingual synsets is discussed together with various ways of reducing their number.

III. MULTILINGUAL SYNSETS
Multilingual synsets are at the core of this project. Naturally emanating from word alignment in parallel corpora, they make a crucial link between semantics in the original bilingual corpora and the development of a WordNet like resource, rich in semantics and semantic relations between words and phrases.

The concept is simple. A synset, as the name suggests, is a set of synonyms. In the context of this paper, its the aligned
words-phrases in the parallel corpora, put together in the form of 4-tuples.

Figure 1 gives a few examples of the synsets. As can be seen many synsets are phrases rather than words. In the example one synset is comprised of four words “shall do so gladly”.

![Fig. 1. Examples of Synsets.](image)

Multilingual synsets help in disambiguating the senses of a word. Translating the English word ‘bank’ with the French ‘banque’ suggests two possible meanings: a financial institution or a collection of a particular kind (e.g., a blood bank), as these words share both meanings, but eliminating the English meaning of a ‘river bank’. Increasing the number of languages could gradually remove all ambiguity, as in the case of \{EN: bank, FR: banque, NL: bank\}. Insofar these lists of words specify a single semantic concept, they can be seen as WordNet-like synsets that makes use of words of several languages, rather than just one. The greater the number of translations in this multilingual WordNet, the clearer the meaning, yet, one might object, the fewer the number of such polyglots, who could benefit from such translations. However, these multilingual synsets can also be useful in a monolingual context, as unique indices that distinguish the individual meanings of a word.

When annotating parallel corpora with lexical semantics, the multilingual synsets become the sense tags and the parallel corpora are tagged with corresponding tags in a single unsupervised process. The idea is as simple as it is elegant: assuming we have a word-aligned parallel corpus with \(n\) languages, annotate each word with a lexical semantic tag consisting of the \(n\)-tuple of aligned words. As a result, all occurrences of a given word in the text for language \(L\) are considered as having the same sense, provided they correspond to (are tagged with) the same multilingual synset.

Two great advantages of this scheme are that it is completely unsupervised, and the fact that, unlike manually tagged corpora using WordNet, all words in the corpus are guaranteed to have a corresponding multilingual synset.

**IV. SYNSET GENERATION AND WSD**

In order to generate the synsets we needed the word-aligned corpora. The Europarl corpus was taken. It was pre-processed, which included among other steps, tokenization of text, lowercasing, removal of empty lines and the removal of XML-tags. After pre-processing a paragraph aligned parallel corpus was obtained. English corpus was used as the pivotal one. All these were fed to GIZA++\(^2\), a standard and freely available tool for word alignment. For alignment, pair-wise corpora were fed into GIZA++ (German with English, French with English, and Greek with English). Thus the output of GIZA++ were pair-wise aligned parallel corpora with markings indicating which words in the target language aligned with which words in English. It might be the case that one word in one language aligns with more than one words in another or it aligns with nothing. Only the aligned words were of any use while generating synsets from the aligned corpora.

For actual synset generation from the aligned corpora we designed our algorithm, which links two or more words in one language together if they align with the same word in another language. The process had to be carried out simultaneously for all the four languages, so as no useful information is lost.

The algorithm links the words of the pivotal language (PL) into phrases and maps all words of the non-pivotal languages to one of these phrases. The array \(a[1..N]\) serves to store in the field \(a[i]\) the number of the phrase to which word \(i\) in the pivotal language belongs. Initially, all PL words are assumed to belong to different phrases (i.e., they form a phrase on their own). Two or more PL words \(a[j], ... a[j+k]\) are placed in the same group if there is a word in another language, which is aligned with all of them. This information is stored by assigning the same phrase number to \(a[j], ... a[j+k]\). The array \(t\) is used to store information about the word alignment between each non-PL and the PL. The assignment \(t[Li]\) represents the fact that the \(i\)-th word in non-PL \(l\) was aligned with the \(k\)-th word in the PL.

Subsequently, each synset is spelt out by producing a phrase in the pivotal language (consisting of one or more PL words with the same phrase number) and extracting for each non-PL all the words that point to a PL word in that group: this final step is straightforward, and due to space limitations is not shown in Figure 2.

While performing the task of synset generation WSD of the original corpus in English was done automatically. That was achieved because the start of each separate phrase in English is numbered with the index number of the first word in that phrase in the whole original corpus. Thus the phrase “shall do so gladly” (reference Fig. 1) is assigned the number 41, which is the index of the word pleasant in the whole original English corpus. Thus the start of each phrase in the English corpus has been assigned a sense tag (the 4-tuple synset) and it constitutes the WSD part of the process.

Part of Speech (POS) is an extra bit of useful information that can be used for WSD [7], [8]. POS tags of the neighbors of the target word help in narrowing down the meanings of the word. We used Brill Tagger [9] to assign POS tags to individual words in the English phrases in the synsets.

The approach described here produces a large number of what we would call ‘proto-synsets’—for a corpus of more than 1.8 million words, there are more than 1.5 million different such synsets. Their number can be reduced and their composition—brought closer to what one would expect to see in a hand-crafted dictionary in the following two ways:

\[^2\text{http://fjoch.com/GIZA++.html}\]
firstly, through the identification and merger of proto-synsets
only varying in word forms corresponding to the same lexical
entry (e.g., flight-X-Y-Z, flights-X-Y-Z); secondly, through
the merger of proto-synsets in which the differences are
limited to words that are synonyms in the given language
(e.g., car-auto-automobile vs car-auto-voiture). These two
approaches are addressed in the following two sections.

V. EDIT DISTANCES

We need to merge the redundant synsets, based on their
syntax and semantics, since morphemes could be both
inflectional and derivational. In inflectional morphemes
the meaning is not changed. Hence both dog and dogs
have the same meaning and dogs is an inflection of dog. In
derivational morphemes, however, the meaning might change.
Thus unhappy is derived from happy, yet they are antonyms
of each other.

Both inflectional and derivational morphemes need to be
taken care of and corresponding synsets merged in order to
reduce the number of synsets and making the resource more
concise and useful. For inflectional morphology we used theedit distance, for derivational we intent to use synonymy
detection, which is discussed in the next section.

Edit distance measures the minimum number of edit steps
required to convert one string into another [10], [11], [12]. The
only three operations allowed are insertion of a character from
the first string, deletion of a character from the first string, or
substitution/replacement of a character in the first string with
a character in the second string. Thus dogs has an edit distance
of 1 with dog, since only a deletion of ‘s’ would suffice for
conversion. There might be more then one ways to conversion,
hence the minimum edit distance is a more useful measure.

We divided the synsets into two groups. The first group
contained all the synsets with frequency one, based on the
English phrase. The other group contained synsets which
have frequency more than one, based on their English phrase.
Pair-wise edit distances were measured between every two
synsets that shared the English phrase. This information would
be used in future to determine which two synsets should be
merged.

VI. SYNONYMY DETECTION

Synonymy is a relationship between words which makes
them inter-substitutable. Yet [13] says that “natural languages
abhor absolute synonyms just as nature abhors a vacuum.”
Absolute synonymy is rare and restricted mostly to technical
terms [14]. Near-synonyms are of greater significance and are
very similar but not completely inter-substitutable or identical.

According to [15] a common approach to synonymy
detection is distributional similarity. Thus synonyms
words share common contexts, and thus they could be
inter-substituted without changing the context. They showed
that use of multilingual resources for extraction of synonyms
had higher precision and recall as compared to the
monolingual resources.

Turney [16] used PMI-IR (Pointwise Mutual Information
and Information Retrieval) to determine the synonymy
between two words. The algorithm maximizes Pointwise
Mutual Information [17], [18], which in turn is based on
c o-occurrence [19].

We can use the above ideas to detect synonymy
between the words/phrases for a given language, then
merge the multilingual proto-synsets that only vary in
this respect. Similarly, we can apply similarity measures
to 4-tuples, e.g., if the words/phrases in all but one
language are the same, or a number of alternatives for
some languages appear together in several permutations,
e.g., car-auto, car-auto-voiture, automobile-auto-auto,
avtomobile-voiture, we can consider them as synonyms.

VII. CONCLUSION

The value of this approach is in its use of unsupervised
techniques that do not require an annotated corpus. In this way,
all words are guaranteed to be tagged with a synset, which is
not often the case with other approaches. This has been done
on a large dataset with more than 1.8 million words. WSD of
such a large corpus is valuable even if the additional benefits
of the lexical resource produced are not considered.

REFERENCES

infrastructure,” Communications of the ACM, vol. 38, no. 11, pp. 33–38,
1995.


