Abstract
We present an algorithm for verb detection of a language in question in a completely unsupervised manner. First, a shallow parser is applied to identify – amongst others – noun and prepositional phrases. Afterwards, a tag alignment algorithm will reveal fixed points within the structures which turn out to be verbs.

Results of corresponding experiments are given for English and German corpora employing both supervised and unsupervised algorithms for part-of-speech tagging and syntactic parsing.

1 Introduction
Recently, along with the growing amount of available textual data, interest in unsupervised natural language processing (NLP) boosts, too.
Especially companies gradually discover its value for market research, competitor analysis and quality assurance to name just a few. During the last decades, language resources were created for many languages, but some domains have very specialized terminology or even particular grammars and for those, no proper resources exist. Hence, unsupervised approaches need to evolve into the direction of information extraction, which still needs huge manual and costly effort in most cases.

In this paper, we want to introduce an approach for unsupervised verb detection solely relying on unsupervised POS tagging and unsupervised shallow parsing. This algorithm will facilitate deep unsupervised parsing as it can provide useful information about verbs along with argument assignments and thus, it is a crucial step for information extraction from data sources for which no suitable language models exist. According to our knowledge, there is no algorithm approaching the problem of unsupervised verb detection so far.

2 Verb detection
Verbs represent natural language relations. The arguments of a verb can be nominal phrases, prepositional phrases or other nominal or prepositional expressions. These phrases can be detected in an unsupervised manner. Besides approaches to chunking (see (Skut and Brants, 1998)), several shallow parsers exist (e. g. unsuParse, see (Hänig et al., 2008; Hänig, 2010)) which are applicable to extract the aforementioned phrase types.
Since unsupervised parsers do not use any a priori knowledge about language, one drawback exists: phrases are not labeled in a human-readable way (e. g. NP or PP), not even if they induce labeled parse trees (see (Reichart and Rappoport, 2008))

2.1 Tag Sequence Alignment
In order to detect verbs we employ a tag sequence alignment algorithm (TSA) which is independent from POS and phrase labels. First, we use shallow parsing to detect significant phrases containing, amongst others, NPs and PPs. Afterwards, we align different sequences of the resulting phrases and POS tags to each other. We assume that verbs dominate the structure of a sentence decisively and mark fixed points within the sequence while their arguments can be exchanged and moved to different positions. In a more formal way:
A sequence \( s \) of a sentence with length \( n \) is defined as
\[
s = (s_0 \ldots s_{n-1})
\]
where \( s_i \) can be a phrase tag or a POS tag. Hence, the sequence of a simple sentence may look like \( (NP \) VBD \( NP \) PP)\). Each sentence can be described as a sequence of tag groups representing phrases.
Such a sequence may contain only one group (the
whole sentence) or up to n groups where each
group consists of exactly one tag (e.
.g. three
groups: (NP), (VBD) and (NP PP)). To build those
groups, the sequence is split at certain indices. So,
every grouping is defined by a set of separation
indices contained in the power set given in Equ. 2.

\[ PI(n) = P\{0 \ldots n - 2\} \]  

Formally, each of the \(2^{n-1}\) possible groupings is
given by

\[
g(s, I) = ((s_0 \ldots s_i_0) (s_{i_0+1} \ldots s_{i_1}) \ldots (s_{i_{n-1}+1} \ldots s_{n-1}))
\]

where \( I \in PI(n) \) is a sorted set of separation in-
dices between two component groups \(|I| = x\).

The similarity of two groupings is defined as

\[
sim_{seq}(g(s, I), g(t, J)) =
\begin{cases}
|I| \neq |J| : & 0 \\
\#i : g(s, I)_i = g(t, J)_i : & 0 \\
\text{else :} & \\
\frac{1}{|I|} \sum_{i=0}^{|I|-1} sim(g(s, I)_i, g(t, J)_i)
\end{cases}
\]

(4)

First, the number of groups has to be equal in both
groupings, otherwise these groupings are not con-
sidered to be a valid alignment. Second, there has
to be at least one exact match containing only sim-
ple POS tags and no phrases as we want to detect
POS tags being fixed points within the sequences.
If these two conditions are met, we can calculate
the similarity as the average of the context similar-
ities between all corresponding groups of the two
groupings\(^2\). In order to find the alignment between
two sequences s and t holding the highest simi-
larity, we match every possible grouping of s with
every possible grouping of t.

2.2 Detection of verbs in a corpus

Having the possibility to calculate the best align-
ments of tag sequences, we apply this algorithm
to a whole corpus. After POS tagging and shal-
low parsing, all sentences are transformed into
their corresponding sequences. We only regard se-
quences with a minimum support of at least 10 oc-
currences within the corpus.

Iteratively, sequences are aligned to each other
starting with the most frequent sequence which is

\(^2\)We apply the cosine measure.

singly split into its components (e. g. (NP VBZ
PP) is split into ((NP) (VBZ) (PP))). Then – in or-
der of frequency – every sequence is either aligned
to an existing sequence (e. g. (NP VBZ NP PP) is
split into ((NP) (VBZ) (NP PP)) due to high simi-
arity to ((NP) (VBZ) (PP))) or represents a new
sequence which is different to the others. A thresh-
old \( \vartheta \) draws the line between those two possibil-
ties. In the latter case, all subsequences of the se-
quence are tested for high context similarity to al-eady detected verbs. This is done to cover verbal
expressions consisting of more than one compo-
ent, e. g. for modal auxiliaries like (MD VB).

Afterwards the new sequence is split into its com-
pounds like the first tag sequence, except for the
subsequence showing high similarity to verbal ex-
pressions which is put into one group.

After processing the most frequent sequences,
several graph structures containing the aligned
sentences are created (e. g. in Figure 1).

Figure 1: Resulting graph structure of several
aligned tag sequences

The part-of-speech building the fixed point in
the graph (as VBZ in the example) is consid-
ered to occupy a central role within the sequences.
Thus, all parts-of-speech (excluding phrases) in all
alignments holding this property and which are not
contained in the phrases extracted by the shallow
parser will be marked as verbs.

2.2.1 Tag list expansion

As we do not use all sequences (only the ones
matching a certain minimum support) and not all
tag sequences achieve a high similarity to other
ones, not all verbs are detected. Hence, we use
all extracted tags \( T \) to generate a set of words \( W_T \)
consisting of all words which are annotated by one
of those tags. Afterwards, we calculate a relative
score for each tag of the tagset expressing the cov-
erage

\[
cov(tag) = \frac{|\text{words annotated by tag} \cap \text{words} \in W_T|}{|\text{words annotated by tag}|}
\]

(5)

We expand the set of extracted verbs to tags which
are well covered by words which already have
been detected. Every POS tag \( t \) with \( \text{cov}(t) \geq 0.5 \) is considered to contain verbs, too.

3 Evaluation

The proposed algorithm is applied to both supervisedly and unsupervisedly annotated corpora to provide comprehensive results. Both configurations were processed for two languages: English and German. We used the corpora en100k and de100k from Projekt Deutscher Wortschatz (see Quasthoff et al., 2006), each containing 100k sentences. We want to point out, that the supervised setup's purpose is only to verify our theory on high quality prerequisites.

For supervised preprocessing steps, we used the Stanford POS Tagger (see Toutanova and Manning, 2000)) and Stanford Parser (see (Klein and Manning, 2003)). Sentence patterns are created by extraction of all kinds of prepositional phrases and noun phrases.

We applied unsuPOS (see Biemann, 2006)) for unsupervised part-of-speech tagging. Afterwards, we trained a model for unsuParse (see Hanig, 2010)) on these data sets for unsupervised shallow parsing (using only phrases with a significance of at least 10% of the most significant one). In this case, we annotated all phrases found by unsuParse.

In either configurations we applied a threshold of \( \vartheta = 0.8 \) and took all sentence patterns having a frequency of at least 10% of the most frequent one into account.

3.1 Part-of-speech tagsets

Each of the four possible setups relies on a different tagset. As it is very important for interpretation of obtained results, we will shortly introduce those tagsets along with the classes containing verbs.

3.1.1 Penn Tree Tagset

The Penn Tree Tagset (see Santorini, 1990)) is applicable to English data. It contains 45 tags containing 7 tags describing verbs. Table 1 gives a short overview about its tags along with their relative frequencies (amongst all tags containing verbs) in the evaluation data set.

3.1.2 Stuttgart-Tübingen Tagset (STTS)

For German data, the Stuttgart-Tübingen Tagset (see Thiel et al., 1999)) is well established. It contains 54 tags, 12 of them contain verbs (see Table 1).

3.1.3 unsuPOS word classes

Unsupervised induced word classes are not labeled in a comparable way as other tagsets. Hence, we give a short overview over the most frequent classes containing verbs in a descriptive way (see Tables 2 and 3). For English, we apply the MEDLINE-model which has been trained on 34 million sentences, the German-model has been trained on 40 million sentences.

4 Results

We calculated precision and recall scores for the extracted verb classes (see Table 4), the corre-
sponding tag sets are given in Table 5.

|               | Precision | Recall | F-Measure |
|---------------|-----------|--------|-----------|
| **English**   |           |        |           |
| supervised    | 1.000     | 0.553  | 0.712     |
| sup. w/ exp.  | 1.000     | 0.894  | 0.944     |
| unsupervised  | 1.000     | 0.440  | 0.611     |
| **German**    |           |        |           |
| supervised    | 1.000     | 0.789  | 0.882     |
| sup. w/ exp.  | 1.000     | 0.816  | 0.899     |
| unsupervised  | 1.000     | 0.627  | 0.771     |

Table 4: Precision, recall and f-measure values

|               | Verb detection | Expansion |
|---------------|----------------|-----------|
| **English**   |                |           |
| supervised    | VBD VBP        | VBN VB    |
|               | VBZ MD         |           |
|               | 26 478 479     | 112 126 336 |
|               |                | 350       |
| unsupervised  | VVFIN VVINF    | VAINF     |
|               | VAFIN VMFIN    | VVIMP     |
|               | 9 37 42        | 135 142 166 |
|               | 334 380        | 175 230 . . . |
| **German**    |                |           |
| supervised    | VVFIN VVINF    | VAINF     |
|               | VAFIN VMFIN    | VVIMP     |
|               | 9 37 42        | 135 142 166 |
|               | 334 380        | 175 230 . . . |
| unsupervised  | VVFIN VVINF    | VAINF     |
|               | VAFIN VMFIN    | VVIMP     |
|               | 9 37 42        | 135 142 166 |
|               | 334 380        | 175 230 . . . |

Table 5: Extracted POS tags

For both the supervised and unsupervised data sets all extracted parts-of-speech contain verbs only. Regarding the supervised data sets for English and German, TSA detects 55.3% and 78.9% of all verbs, respectively. Tag set expansion yields a significant improvement for English (raising recall to 89.4%), while the improvement for German is marginal. This observation is not very surprising as German is morphologically richer than English.

The results on unsupervised data are perfectly accurate, too. For this setup, tag list expansion does not have a measurable impact on our results (approx. 0.02%) and can be neglected. However, expansion adds some classes including some incorrect ones (the italic ones in Table 5). The lower recall results from a much higher number of different word classes (about 500 in our case) induced by an unsupervised POS tagger. The lack of POS tag disambiguation is the reason for the inefficiency of our expansion step, since almost no word form is tagged by different tags.

5 Conclusions and further work

We have shown that alignment of tag sequences containing chunks or shallow parses can detect verbs in a completely unsupervised manner. Although the actual alignment covers the most common verb classes, expansion increases the number of correctly detected verbs.

In the future, we plan to evaluate other approaches to unsupervised POS tagging. We also want to incorporate unsupervised morphological analysis to improve the performance on morphologically rich languages.

References

Chris Biemann. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In Proceedings of the COLING/ACL-06 Student Research Workshop.

Christian Hänig. 2010. Improvements in Unsupervised Co-Occurrence Based Parsing. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning.

Christian Hänig, Stefan Bordag, and Uwe Quasthoff. 2008. Unsuparse: Unsupervised parsing with unsupervised part of speech tagging. In Proceedings of the Sixth International Language Resources and Evaluation (LREC’08).

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In In Advances in Neural Information Processing Systems 15 (NIPS).

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus portal for search in monolingual corpora. In Proceedings of the LREC 2006.

Roi Reichart and Ari Rappoport. 2008. Unsupervised induction of labeled parse trees by clustering with syntactic features. In Proceedings of the 22nd International Conference on Computational Linguistics.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, University of Pennsylvania.

W. Skut and T. Brants. 1998. Chunk tagger-statistical recognition of noun phrases. Arxiv preprint cmp-lg/9807007.

C. Thielen, A. Schiller, S. Teufel, and C. Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, University of Stuttgart and University of Tübingen.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In In EMNLP/VLC 2000.