A Cascaded Classification Approach to Semantic Head Recognition

Lukas Michelbacher Alok Kothari Martin Forst[†]
Christina Lioma Hinrich Schütze
Institute for NLP
University of Stuttgart
{michells,kotharak,liomaca}@ims.uni-stuttgart.de
[†]Microsoft
martin.forst@microsoft.com

Abstract

Most NLP systems use tokenization as part of preprocessing. Generally, tokenizers are based on simple heuristics and do not recognize multi-word units (MWUs) like *hot dog* or *black hole* unless a precompiled list of MWUs is available. In this paper, we propose a new cascaded model for detecting MWUs of arbitrary length for tokenization, focusing on noun phrases in the physics domain. We adopt a classification approach because – unlike other work on MWUs – tokenization requires a completely automatic approach. We achieve an accuracy of 68% for recognizing non-compositional MWUs and show that our MWU recognizer improves retrieval performance when used as part of an information retrieval system.

1 Introduction

Most NLP systems use tokenization as part of preprocessing. Generally, tokenizers are based on simple heuristics and do not recognize multi-word units (MWUs) like *hot dog* or *black hole*. Our long-term goal is to build MWU-aware tokenizers that are used as part of the standard toolkit for NLP preprocessing alongside part-of-speech and named-entity tagging.

We define an MWU as a sequence of words that has properties that cannot be inferred from the component words (cf. e.g. Manning and Schütze (1999, Ch. 5), Sag et al. (2002)). The most important of these properties is non-compositionality, the fact that the meaning of a phrase cannot be predicted from the meanings of its component words. For example, a *hot dog* is not a hot animal but a sausage in a bun and a *black hole* in astrophysics is a region of space with special properties, not a dark cavity.

The correct recognition of MWUs is an important building block of many NLP tasks. For example, in information retrieval (IR) the query *hot dog* should not retrieve documents that only contain the words *hot* and *dog* individually, outside of the phrase *hot dog*.

In this study, we focus on noun phrases in the physics domain. For specialized domains such as physics, adaptable and reliable MWU recognition is of particular importance because comprehensive and up-to-date lists of MWUs are not available and would have to be created by hand. We chose noun phrases because domain-specific terminology is commonly encoded in noun phrase MWUs; other types of phrases – e.g., verb constructions – rarely give rise to fixed domain-specific multi-word sequences that should be treated as a unit.

We cast the task of MWU tokenization as *semantic head recognition* in this paper. The importance of syntactic heads for many NLP tasks is generally accepted. For example, in coreference resolution identity of syntactic heads is predictive of coreference; in parse disambiguation, the syntactic head of a noun phrase is a powerful feature for resolving attachment ambiguities. However, in all of these cases, the syntactic head is only an approximation of the information that is really needed; the underlying assumption made when using the syntactic head as a substitute for the entire phrase is that the syntactic head is representative of the phrase. This is not the case when the phrase is non-compositional.

We define the *semantic head* of a noun phrase as the non-compositional part of a phrase. Semantic heads would serve most NLP tasks better than syntactic heads. For example, a coreference resolution system is misled if it looks at syntactic heads to de-

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 793–803,
Edinburgh, Scotland, UK, July 27–31, 2011.

termine possible coreference of *a hot dog ... the dog* in *I first ate a hot dog and then fed the dog*. This is not the case for a system that makes the decision based on the semantic heads *hot dog* of *a hot dog* and *dog* of *the dog*.

The specific NLP application we evaluate in this paper is information retrieval. We will show that semantic head recognition improves the performance of an information retrieval system.

We introduce a cascaded classification framework for recognizing semantic heads that allows us to treat noun phrases of arbitrary length. We use a number of previously proposed features for recognizing non-compositionality and semantic heads. In addition, we compare three features that measure contextual similarity.

Our main contributions in this paper are as follows. First, we introduce the notion of semantic head, in analogy to syntactic head, and propose semantic head recognition as a new component of NLP preprocessing. Second, we develop a cascaded classification framework for semantic head recognition. Third, we investigate the utility of contextual similarity for detecting non-compositionality and show that it significantly enhances a baseline semantic head recognizer. However, we also identify a number of challenges of using contextual similarity in high-confidence semantic head recognition. Fourth, we show that our approach to semantic head recognition improves the performance of an IR system.

Section 2 discusses previous work. In Section 3 we introduce semantic heads and present our cascaded model for semantic head recognition. In Section 4, we describe our data and three different measures of contextual similarity. Section 5 introduces the classifier and its features. Section 6 presents classification results and discussion. Section 7 describes the information retrieval experiments. In Section 8 we present our conclusions.

2 Related Work

While there is a large number of publications on MWUs and collocation extraction, the general problem of automatic MWU detection for the specific purpose of tokenization has not been investigated before to our knowledge.

The classic approach to identifying collocations and MWUs is to apply statistical association measures (AMs) to n-grams extracted from a corpus – often combined with various linguistic heuristics and other filters, resulting in candidate lists. Choueka (1988) and the XTRACT system (Smadja, 1993) are well-known examples of this approach.

More recent approaches such as Pecina (2010) and Ramisch et al. (2010) combine classifiers with association measures. Although our approach is classification-based as well, our data set has a more realistic size than Pecina (2010)'s (1 billion words vs 1.5 million words) and we work on noun phrases of arbitrary length (instead of just bigrams). The `mwetoolkit`[1] by Ramisch et al. (2010) aims to be a software package for lexicographers and its features are limited to a small set of association measures that do not consider marginal frequencies. Neither of these two studies includes evaluation in the context of an application.

Lin (1999) defines a decision criterion for non-compositional phrases based on the change in the mutual information of a phrase when substituting one word for a similar one based on an automatically constructed thesaurus. The method reaches 15.7% precision and 13.7% recall.

In terms of the extraction of domain-specific MWUs, cross-language methods have been proposed that make use of the fact that an MWU in one language might be expressed as a single word in another. Caseli et al. (2009) utilize word alignments in a parallel corpus; Attia et al. (2010) exploit the links between article names of different-language Wikipedias to search for many-to-one translations. We did not pursue a cross-language approach because we strive for a self-contained method of MWU recognition that operates on a single textual resource.

Non-compositionality and distributional semantics. In recent years, a number of studies have investigated the relationship between distributional semantics and non-compositionality. These studies compute the similarity between words and phrases represented as semantic vectors in a word space model. A semantic vector of a word is the accumulation of the particular contexts in which the word

[1] `http://sourceforge.net/projects/mwetoolkit/`

appears. The underlying idea is similar to Lin's: the meaning of a non-compositional phrase somehow deviates from what one would expect given the semantic vectors of parts of the phrase. The standard measure to compare semantic vectors is cosine similarity. The questions that arise are (i) which vectors to compare, (ii) how to combine the vectors of the parts and (iii) from what point on a certain dissimilarity indicates non-compositionality. To our knowledge, there are no generally accepted answers to these questions.

Regarding (i), Schone and Jurafsky (2001) compare the semantic vector of a phrase p and the vectors of its component words in two ways: one includes the contexts of p in the construction of the semantic vectors of the parts and one does not. Regarding (ii), they suggest weighted or unweighted sums of the semantic vectors of the parts.

Baldwin et al. (2003) investigate semantic decomposability of noun-noun compounds and verb constructions. They address (i) by comparing the semantic vectors of phrases with the vectors of their parts *individually* to detect meaning changes; e.g., they compare *vice president* to *vice* and *president*.

We propose a new method that compares phrases with their alternative phrases, in the spirit of Lin (1999)'s substitution approach (see Section 4.3). Our rationale is that context features should be based on contexts that are syntactically similar to the phrase in question.

With respect to (iii), the above-mentioned studies use ad hoc thresholds to separate compositional and non-compositional phrases but do not offer a principled decision criterion.[2] In contrast, we train a statistical classifier to learn a decision criterion.

There is a larger body of work concerning non-compositionality which revolves around the problem of literal (compositional) vs. non-literal (non-compositional) usage of idiomatic verb constructions like *to break the ice* or *to spill the beans*. Some studies approach the problem with semantic vector comparisons in the style of Schone and Jurafsky (2001), e.g Katz and Giesbrecht (2006) and Cook et al. (2007). Other approaches use word-alignment (e.g. Moirón and Tiedemann (2006)) or a combination of heuristic and linguistic features (e.g. Diab and Bhutada (2009), Li and Sporleder (2010)). Even though there is some methodological overlap between our approach and some of the verb-oriented studies, we believe that verb constructions have properties that are quite different from noun phrases. For example, our definition of alternative vector relies on the fact that most noun phrase MWUs are fixed and exhibit no syntactic variability. In contrast, verb constructions are often discontinuous.

The motivation for most work on MWU detection is lexicography, terminology extraction or the creation of machine-readable dictionaries. Our motivation – tokenization in a preprocessing setting – is different from this earlier work.

3 Semantic Heads and Cascaded Model

We cast the task of MWU tokenization as *semantic head recognition* in this paper. We define the semantic head of a noun phrase as the *largest non-compositional part of the phrase that contains the syntactic head*. For example, *black hole* is the semantic head of *unusual black hole* and *afterglow* is the semantic head of *bright optical afterglow*; in the latter case syntactic and semantic heads coincide.

Semantic heads would serve most NLP tasks better than syntactic heads. The attachment ambiguity of the last noun phrase in *he bought the hot dogs in a packet* can be easily resolved for the semantic head *hot dogs* (food is often in a packet), but not as easily for the syntactic head *dogs* (dogs are usually not in packets). Indeed, we will show in Section 7 that semantic head recognition improves the performance of an IR system.

The semantic head is either a single noun or a non-compositional noun phrase. In the latter case, the modifier(s) introduce(s) a non-compositional, unpredictable shift of meaning; *hot* shifts the meaning of *dog* from live animal to food. In contrast, the compositional meaning shift caused by *small* in *small dog* is transparent. The semantic head always contains the syntactic head; for compositional phrases, syntactic head and semantic head are identical.

To determine the semantic head of a phrase, we use a cascaded classification approach. The cascade

[2] Lin (1999) uses a well-defined criterion but his approach is not based on vector similarity.

(1)		*neutron*	**star**	
(2)	*unusual*	black	**hole**	
(3)	*bright*	optical	**afterglow**	
(4)		*small*	**moment**	of inertia

Figure 1: Example phrases with modifiers. Peripheral elements are set in italics, syntactic heads in bold.

comes into play in all aspects of our study: the rating experiments with human subjects, data extraction, feature design and classification itself.

We need a cascade because we want to recognize the semantic head in noun phrases of arbitrary length. The starting point is a phrase of length n: $p = w_1 \dots w_n$. We distinguish between the syntactic head of a phrase and the remaining words, the modifiers. Figure 1 shows phrases of varying syntactic complexity. The syntactic head is marked in bold. The model accommodates pre-nominal modifiers as in examples (1) through (3) and post-nominal modifiers like PPs in example (4).

Among the modifiers, there is a distinguished element, the *peripheral element* u (italicized in the examples). The remaining words are called the *rest* v. We can now represent any phrase p as $p = uv$.[3] The element u is always the outermost modifier. *of*-PPs are treated as a single modifier and they take precedence over pre-nominal modification because this analysis is dominant in our gold standard data. This means that in the phrase *small moment of inertia*, *small* (and not *of inertia*) is the peripheral element u.

Cascaded classification then operates as shown in Figure 2. In each iteration, the classifier decides whether the relation between the current peripheral element u and the rest v is compositional (C) or non-compositional (NC). If the relation is NC, processing stops and uv is returned as the semantic head of p. If the relation is compositional, u is discarded and classification continues with v as the new input phrase, which again is represented in the form $u'v'$. In case there is no more peripheral element u, i.e. the new v is a single word, it is returned as the semantic head of p.

Table 1 shows two examples. For the fully compositional phrase *bright optical afterglow*, the process runs all the way down to the syntactic head *afterglow* which is also the semantic head. In the second case, the process stops earlier, in step 2, because the classifier finds that the relation between *moment* and *of inertia* is NC. This means that the semantic head of *small moment of inertia* is *moment of inertia*.

[3] We use the abstract representation $p = uv$ even though u can appear after v in the surface form of p.

function $recognize\_semantic\_head(p)$
  $u \leftarrow peripheral(p)$
  $v \leftarrow rest(p)$
  while $decision(u, v) \neq NC$ **do**
    $u \leftarrow peripheral(v)$
    if $u = \emptyset$ **then**
      return v
    $v \leftarrow rest(v)$
  return uv

Figure 2: Cascaded classification of p

step	u	v	decision
1	*bright*	optical **afterglow**	C
2	*optical*	**afterglow**	C
3	$\emptyset$	**afterglow**	
1	*small*	**moment** of inertia	C
2	*of inertia*	**moment**	NC

Table 1: Cascaded decision processes

4 Corpus and Feature Definitions

4.1 Candidate phrases

As our corpus, we use the iSearch collection, a one billion word collection of documents from the physics domain (Lykke et al., 2010). We tokenized the collection by splitting on white space and adding sentence boundaries and part-of-speech tags to the output. With part-of-speech information, the identification of MWU candidates is easy, fast and reliable.

We extracted all noun phrases from the collection that consist of a head noun with up to four modifiers – almost all domain-specific terminology in our collection is captured by this pattern. The pre-nominal modifiers can be nouns, proper nouns, adjectives or cardinal numbers.

The baseline accuracy of a classifier that always chooses compositionality is very high ($> 90\%$) for

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Table 2: 2-by-2 contingency tables with observed and marginal frequencies

phrases of the type *[noun] of the/a [noun] (sg.)* (e.g. *rest of the paper*) and *[noun] of [noun] (pl.)* (e.g. *series of papers*). We therefore restrict postnominal modifiers to prepositional phrases with the word *of* followed by a non-modified, indefinite, singular noun, e.g., *speed of light* or *moment of inertia*.

Out of all phrases extracted with part-of-speech patterns, we keep only the ones that appear more often than 50 times because it is hard to compute reliable features for less frequent phrases. All experiments were carried out with lemmatized word forms. We refer to lemmas as words if not noted otherwise.

4.2 Association measures

Statistical association measures are frequently used for MWU detection and collocation extraction (e.g. Schone and Jurafsky (2001), Evert and Krenn (2001), Pecina (2010)).

We use all measures used by Schone and Jurafsky (2001) that can be derived from a phrase's contingency table. These measures are Student's t-score, z-score, χ^2, pointwise mutual information (MI), Dice coefficient, frequency, log-likelihood (G^2) and symmetric conditional probability.

We define the AMs in Table 3 based on the notation for the contingency table shown in Table 2 (cf. Evert (2004)). O_{ij} is observed frequency and $E_{ij} = \frac{R_i C_j}{N}$ expected frequency.

The AMs are designed to deal with two random variables U and V that traditionally represent single words. In our model, we use U to represent peripheral elements u and V for rests v.

association measure	formula
student's t-score (am_t)	$\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$
z-score (am_z)	$\frac{O_{11}-E_{11}}{\sqrt{E_{11}}}$
chi-square (am_{χ^2})	$\sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
pointwise mutual information (am_{MI})	$\log \frac{O_{11}}{E_{11}}$
Dice coefficient (am_D)	$\frac{2O_{11}}{R_1+C_1}$
frequency (am_f)	O_{11}
log-likelihood (am_{G^2})	$2\sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}$
symmetric conditional probability (am_{scp})	$\frac{{O_{11}}^2}{R_1 C_1}$

Table 3: Association measures

4.3 Word space model

As our baseline, we use two methods of comparing semantic vectors: *sj1* and *sj2*, both introduced by Schone and Jurafsky (2001). They experimented with variants of sj1 and sj2, but found no large differences. In addition, we introduce our own approach *alt*.

Method sj1 compares the semantic vector of a phrase p with the sum of the vectors of its parts. Method sj2 is like sj1, except the contexts of p are not part of the semantic vectors of the parts. Method alt compares the semantic vector of a phrase with its alternative vector. In the definitions below, s represents a vector similarity measure, $w(p)$ a general semantic vector of a phrase p and $w^*(w_i)$ the semantic vector of a part w_i of a phrase p that does not include the contexts of occurrences of w_i that were part of p itself.

sj1 $s(w(\textit{black hole}),\ w(\textit{black}) + w(\textit{hole}))$

sj2 $s(w(\textit{black hole}),\ w^*(\textit{black}) + w^*(\textit{hole}))$

alt $s(w(\textit{black hole}), \sum_u w(u,\ \textit{hole}));\ u \neq \textit{black}$

For the third comparison, we build the *alternative vector* as follows. For a phrase $p = uv$ with peripheral element u and rest v, we call the phrase

$p' = u'v$ an *alternative phrase* if the rest v is the same and $u' \neq u$. E.g., *giant star* is an alternative phrase of *neutron star* and *isolated neutron star* is an alternative of *young neutron star*. The alternative vector of p is then the semantic vector that is computed from the contexts of all of p's alternative phrases. The alternative vector is a representation of the contexts of v except for those modified by u. This technique bears resemblance to the substitution approach of Lin (1999). The difference is that he relies on a similarity thesaurus for substitution and monitors the change in mutual information for each substitution individually whereas we substitute with general alternative modifiers and combine the alternative contexts into one vector for comparison.

Previous work has compared the semantic vector of a phrase with the vectors of its components. Our approach is more "head-centric" and only compares phrases in the same syntactic configuration. Our question is: Is the typical context of the head *hole* if it occurs with a modifier that is not *black* different from when it occurs with the modifier *black*?

We used a bag-of-words model and a window of ± 10 words for contexts to create semantic vectors. We only kept the content words in the window which we defined as words that are tagged as either a noun, verb, adjective or adverb. To add information about the variability of syntactic contexts in which phrases occur, we add the words immediately before and after the phrase with positional markers (-1 and $+1$, respectively) to the vector. These words were not subject to the content-word filter. The dimensionality of the vectors is then $3V$ where V is the size of the vocabulary: V dimensions each for bag-of-words, left and right syntactic contexts. We did not include vectors for the stop word *of* for sj1 and sj2.

4.4 Non-compositionality judgments

Since the domain of the corpus is physics, highly specialized vocabulary had to be judged. We employed domain experts as raters (one engineering and two physics graduate students).

In line with the cascaded model, the raters where asked to identify the semantic head of each candidate phrase. If at least two raters agreed on a semantic head of a phrase we made this choice the semantic head in the gold standard. The final gold standard comprises 1560 phrases.

We computed raw agreement of each rater with the gold standard as the percentage of correctly recognized semantic heads – this is the task that the classifier addresses. Agreement is quite high at 86.5%, 88.3% and 88.5% for the three raters. In addition, we calculated chance-corrected agreement with Cohen's κ on the first decision task against the gold standard (see Section 6). As expected, agreement decreases, but is still substantial at 74.0%, 78.2% and 71.8% for the three raters.

5 Classifier

We use the Stanford maximum entropy classifier for our experiment.[4] We randomly split the data into a training set of 1300 and a held-out test set of 260 pairs.

We use the eight AMs and the cosine similarities sim_{sj1}, sim_{sj2} and sim_{alt} described in Section 4.3 as features for the classifier. Cosine similarity should be small if a phrase is non-compositional and large if it is compositional. In other words, if the contexts of the candidate phrase are too dissimilar to the contexts of the sum of its parts or to the alternative phrases, then we suspect non-compositionality.

Feature values are binned into 5 bins. We applied a log transformation to the four AMs with large values: am_f, am_{G^2}, am_{χ^2} and am_z. For our application there is little difference between statistical significance at $p < .001$ and $p < .00001$. The log transformation reduces the large gap in magnitude between high significance and very high significance. If co-occurrence of u and v in uv is below chance, then we set the association scores to 0 since this is an indication of compositionality (even if it is highly significant).

Since AMs have been shown to be correlated (e.g. Pecina (2010)), we first perform feature selection on the AM features. We tested accuracy of all $2^r - 1$ non-empty combinations of the $r = 8$ AM features on the task of deciding whether the first decision during the classification of a phrase was C or NC. We then selected those AM features that were part of at least one top 10 result in each fold. Those features were am_t, am_f and am_{scp}.

The main experiment combines these three se-

[4] http://nlp.stanford.edu/software/classifier.shtml

lected AM features with all possible subsets of context features. We train on the 1300-element training set and test on the 260-element test set.

6 Results and Discussion

We ran three evaluation modes: *dec-1st*, *dec-all*, and *semh*. Mode *dec-1st* only evaluates the first decision for each phrase; the baseline in this case is .554 since 55.4% of the first decisions are C. In mode *dec-all*, we evaluate all decisions that were made in the course of recognizing the semantic head. This mode emphasizes the correct recognition of semantic heads in phrases where multiple correct decisions in a row are necessary. We define the confidence for multi-decision classification as the product of the confidence values of all intermediate decisions. There is no obvious baseline for *dec-all* because the number of decisions depends on the classifier – a classifier whose first decision on a four-word phrase is NC makes one decision, another one may make three. The mode *semh* evaluates how many semantic heads were recognized correctly. This mode directly evaluates the task of semantic head recognition. The baseline for *semh* is the tokenizer that always returns the syntactic head; this baseline is .488.[5] Table 4 shows 8×3 runs, corresponding to the three modes tested on the AM features (am_t, am_f, and am_{scp}) and the eight possible subsets of the three context features.

For all modes, the best result is achieved with base AMs combined with the sim_{alt} feature; the accuracies are .692, .703 and .680. The improvements over the baselines (for *dec-1st* and *semh*) are statistically significant at $p < .01$ (binomial test, $n = 260$).

For *semh*, accuracy without any context features is .603; this is significantly better than the .488 baseline ($p < .01$). Performance with only the base AM features is significantly lower than the best context feature experiment (.680) at $p < .01$ and significantly lower than the worst context feature experiment (.653) at $p < .1$. However, the differences between the context feature runs are not significant.

When the semantic head recognizer processes a phrase, there are four possible results. Result r_{semh}: the semantic head is correctly recognized and it is distinct from the syntactic head. Result r_{synth}: the semantic head is correctly recognized and it is identical to the syntactic head. Result r_+: the semantic head is not correctly recognized because the cascade was stopped too early, i.e., a compositional modifier that should have been removed was kept. Result r_-: the semantic head is not correctly recognized because the cascade was stopped too late, i.e., a modifier causing a non-compositional meaning shift was removed. Table 5 shows the distribution of result types. It shows that r_+ is the more common error: the classifier more often regards compositional relations as non-compositional than vice versa.

type	freq	definition
r_{semh}	92	sem. head correct ($\neq$ synt. head)
r_{synth}	85	sem. head correct (= synt. head)
r_+	48	sem. head too long
r_-	35	sem. head too short
all	260	

Table 5: Distribution of result types

Table 6 shows the top 20 classifications where the semantic head was not the same as the syntactic head sorted by confidence in descending order. In the third column "phrase ..." we list the candidates with semantic heads in bold. The columns to the right show the predicted semantic head and the feature values. All five errors in the list are of type r^+.

Two r^+ phrases are *schematic view* and *many others*. The two phrases are clearly compositional and the classifier failed even though the context feature points in the direction of compositionality with a value greater than .5. It can be argued that *many others* is a trivial example that does not require complex machinery to be identified as compositional, e.g. by using a stop list. We included it in the analysis since we want to be able to process arbitrary phrases without additional hand-crafted resources.

Another incorrect classification occurs with the phrase *massive star birth*[6] for which *star birth* was annotated as the semantic head. Here we have a case where the peripheral element *massive* does not mod-

[5] The baseline could be improved with simple heuristics, e.g. "uv contains capital letter" $\rightarrow NC$. However, this feature only results in a 2% improvement compared to the baseline.

[6] i.e. the birth of a massive star, a certain type of star with very high mass

mode	baseline	context feature	context feature subsets							
		sim_{alt}	-	•	•	•	•	-	-	-
		sim_{sj1}	-	-	•	–	•	•	-	•
		sim_{sj2}	-	-	-	•	•	-	•	•
dec-1st	.554		.604	.692	.669	.685	.677	.654	.654	.662
dec-all	-		.615	.703	.681	.696	.688	.666	.669	.675
semh	.488		.603	**.680**	.657	.673	.665	.653	.653	.661

Table 4: Performance for base AM features plus context feature subsets. A '•' indicates the use of the corresponding context feature.

ify the syntactic head *birth* but *massive star* is itself a complex modifier. In the test set, 5% of the phrases exhibit structural ambiguities of this type. Our system cannot currently deal with this phenomenon.

The remaining r^+ phrases are *peculiar velocity* and *local group*. However, Wikipedia lists both phrases with an individual entry defining the former as *the true velocity of an object, relative to a rest frame*[7] and the latter as *the group of galaxies that includes Earth's galaxy, the Milky Way*[8]. Both definitions provide evidence for non-compositionality since the velocity is not peculiar (as in strange) and the scope of *local* is not clear without further knowledge. Arguably, in these cases our method chose a justifiable semantic head, but the raters disagreed.[9]

For NLP preprocessing, it is acceptable to sacrifice recall and only make high-confidence decisions on semantic heads. A tokenizer that reliably detects a subset of MWUs is better than one that recognizes none. However, our attempts to use the sim_{alt} recognizer (bold in Table 4) in this way were not successful. Precision is .680 for confidence $>$ .7 and does not exceed .770 for higher confidence values.

To understand this effect, we analyzed the distribution of sim_{alt} scores. Surprisingly, moderate similarity between .4 and .6 is a more reliable indicator for NC than low similarity $<$ .3. Our intuition for using distributional semantics in Section 2 was that low similarity indicates non-compositionality. This does not seem to hold for the lowest similarity values possibly because they are often extreme cases in terms of distribution and frequency and then give rise to unreliable decisions. This means that the context features enhance the overall performance of the classifier, but they are unreliable and do not support the high-confidence decisions we need in NLP preprocessing.

For comparison, the classifier that only uses AM features achieves 90% precision at 14% recall with confidence $>$ .7 – although it has lower overall accuracy than the sim_{alt} recognizer. We are still in the process of analyzing these results and decided to use the AM-only recognizer for the IR experiment because it has more predictable performance.

In summary, the results show that, for the recognition of semantic heads, basic AMs offer a significant improvement over the baseline. We have shown that some wrong decisions are defensible even though the gold standard data suggests otherwise. Context features further increase performance significantly, but surprisingly, they are not of clear benefit for a high-confidence classifier that is targeted towards recognizing a smaller subset of semantic heads with high confidence.

7 Information Retrieval Experiment

Typically, IR systems do not process non-compositional phrases as one semantic entity, missing out on potentially important information captured by non-compositionality. This section illustrates one way of adjusting the retrieval process so that non-compositional phrases are processed as semantic entities that may enhance retrieval performance. The underlying hypothesis is that, given

[7] http://en.wikipedia.org/wiki/Peculiar_velocity

[8] http://en.wikipedia.org/wiki/Local_group

[9] Further evidence that *local group* is non-compositional is the fact that one of the domain experts annotated the phrase as non-compositional but was overruled by the other two.

c.	type	phrase (semantic head in bold)	predicted semantic head	am_t	am_f	am_{cp}	sim_{alt}
.99	r_{semh}	**ellipsoidal figure of equilibrium**	ellipsoidal figure of equilibrium	18.03	325	6.23e-01	.219
.99	r_{semh}	**point spread function**	point spread function	95.03	9056	2.33e-01	.529
.99	r_{+}	massive **star birth**	massive star birth	19.99	402	4.81e-03	.134
.98	r_{semh}	**high angular resolution imaging**	high angular resolution imaging	13.07	179	1.27e-03	.173
.98	r_{semh}	**integral field spectrograph**	integral field spectrograph	24.20	586	4.12e-02	.279
.98	r_{+}	local **group**	local group	153.54	24759	8.73e-03	.650
.98	r_{semh}	**neutral kaon system**	neutral kaon system	1.38	108	4.17e-03	.171
.97	r_{semh}	**IRAF task**	IRAF task	49.07	2411	2.96e-02	.517
.92	r_{semh}	**easy axis**	easy axis	44.66	2019	2.79e-03	.599
.89	r_{+}	schematic **view**	schematic view	40.56	1651	8.06e-03	.612
.87	r_{semh}	**differential resistance**	differential resistance	31.71	1034	6.38e-04	.548
.86	r_{semh}	**TiO band**	TiO band	36.84	1372	2.21e-03	.581
.86	r_{+}	many **others**	many others	97.76	9806	6.54e-03	.708
.86	r_{semh}	**VLBA observation**	VLBA observation	43.95	2004	9.35e-04	.648
.85	r_{+}	peculiar **velocity**	peculiar velocity	167.63	28689	2.37e-02	.800
.84	r_{semh}	**computation time**	computation time	43.80	1967	1.35e-03	.657
.83	r_{semh}	**Land factor**	Land factor	21.15	453	6.30e-04	.360
.83	r_{semh}	**interference filter**	interference filter	31.44	1002	1.27e-03	.574
.83	r_{semh}	**line formation calculations**	line formation calculations	14.20	203	1.96e-03	.381
.82	r_{semh}	**Wess-Zumino-Witten term**	Wess-Zumino-Witten term	9.60	94	8.12e-05	.291

Table 6: The 20 most confident classifications where the prediction is semantic head $\neq$ syntactic head. "c." = confidence

a query that contains a non-compositional phrase, boosting the retrieval weight of documents that contain this phrase will improve overall retrieval performance.

We do this boosting using Indri's[10] combination of the language modeling and inference network approaches (Metzler and Croft, 2004), which allows assigning different degrees of belief to different parts of the query. This belief can be drawn from any suitable external evidence of relevance. In our case, this source of evidence is the knowledge that certain query terms constitute a non-compositional phrase. Under this approach, and using the *#weight* and *#combine* operators for combining beliefs, the relevance of a document D to a query Q is computed as the probability that D generates Q, $P(Q|D)$:

$$P(Q|D) = \prod_{t \in Q} P(t|D)^{\frac{w_t}{W}} \qquad (W = \sum_{t \in Q} w_t) \quad (1)$$

where t is a term and w_t is the belief weight assigned to t. The higher w_t is, the higher the rank of documents containing t. In this work, we distinguish between two types of query terms: terms occurring in non-compositional phrases (Q_{nc}), and the remaining query terms (Q_c). Terms $t \in Q_{nc}$ receive belief weight w_{nc} and terms $t \in Q_c$ belief weight w_c, ($w_{nc} + w_c = 1$ and $w_{nc}, w_c \in [0, 1]$). To boost the ranking of documents containing non-compositional phrases, we increase w_{nc} at the expense of w_c. We estimate $P(t|D)$ in Eq. 1 using Dirichlet smoothing (Zhai and Lafferty, 2002).

We use Indri for indexing and retrieval without removing stopwords or stemming. This choice is motivated by two reasons: (i) We do not have a domain-specific stopword list or stemmer. (ii) Baseline performance is higher when keeping stopwords and without stemming, rather than without stopwords and with stemming.

We use the iSearch collection discussed in Section 4. It comprises 453,254 documents and a set of 65 queries with relevance assessments. To match documents to queries without any treatment of non-compositionality (baseline run), we use the Kullback-Leibler language model with Dirichlet smoothing (KL-Dir) (Zhai and Lafferty, 2002). We applied the preprocessing described

[10] `http://www.lemurproject.org/`

run	MAP	REC	P20
baseline	0.0663	770	0.1385
real NC	**0.0718**	**844**	**0.1538**
pseudo NC_1	0.0664	788	0.1385
pseudo NC_2	0.0658	782	0.1462
pseudo NC_3	0.0671	777	0.1477
pseudo NC_4	0.0681	807	0.1462
pseudo NC_5	0.0670	783	0.1423

Table 7: IR performance without considering non-compositionality (*baseline*), versus boosting real and pseudo non-compositionality (*real NC, pseudo* NC_i).

in Section 4 to the queries and identified non-compositional phrases with the base AM classifier from Section 5. Our approach for boosting the weight of these non-compositional phrases uses the same retrieval model enhanced with belief weights as described in Eq. 1 (real NC run). In addition, we include five runs that boost the weight of pseudo non-compositional phrases that were created randomly from the query text (pseudo NC runs). These pseudo non-compositional phrases have exactly the same length as the observed non-compositional phrases for each query. We measure retrieval performance in terms of mean average precision (MAP), precision at 20 (P20), and recall (REC, number of relevant documents retrieved – total is 2878). For each evaluation measure separately, we tune the following parameters and report the best performance: (i) the smoothing parameter μ of the KL-Dir retrieval model ($\mu \in \{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000\}$, following Zhai and Lafferty (2002)); (ii) the belief weights $w_{nc}, w_c \in \{0.1, \ldots, 0.9\}$ in steps of 0.1 while preserving $w_{nc} + w_c = 1$ at all times.

Table 7 displays retrieval performance of our approach against the baseline and five runs with pseudo non-compositional phrases. We see a 9.61% improvement in the number of relevant retrieved documents over the baseline. MAP and P20 also show improvements. Our approach is better than any of the 5 random runs on all three metrics – the probability of getting such a good result by chance is $\frac{1}{2^5} < .05$, and thus the improvements are statistically significant. On doing a query-wise analysis of MAP scores, we find that large improvements over the baseline occur when a non-compositional phrase aligns with what the user is looking for. The system seems to retrieve more relevant documents in that case. E.g., the improvement in MAP is 0.0977 for query #19. The user was looking for *"articles ... on making tunable vertical cavity surface emitting laser diodes"* and *laser diodes* was one of the non-compositional phrases recognized. On the other hand, a decrease in MAP occurs for non-compositional phrases unrelated to the information need. In query #4 the user is looking for *"protein-protein interaction, the surface charge distribution of these proteins and how this has been investigated with Electrostatic Force Microscopy"* and though non-compositional phrases such as *Force Microscopy* are recognized, these do not reflect the core information need *"The proteins of interest are the Avidin-Biotin and IgG-anti-IgG systems"*.

8 Conclusion

We have presented an approach to improving tokenization in NLP preprocessing that is based on the notion of semantic head. Semantic heads are – in analogy to syntactic heads – the core meaning units of phrases that cannot be further semantically decomposed. To perform semantic head recognition for tokenization, we defined a novel cascaded model and implemented it as a statistical classifier that used previously proposed and new context features. We have shown that the classifier significantly outperforms the baseline and that context features increase performance. We reached an accuracy of 68% and argued that even a semantic head recognizer restricted to high-confidence decisions is useful – because reliably recognizing a subset of semantic heads is better than recognizing none. We showed that context features increase the accuracy of the classifier, but undermine the confidence assessments of the classifier, a result we are still analyzing. Finally, we showed that even in its preliminary current form the semantic head recognizer is able to improve the performance of an IR system.

Acknowledgments

This work was funded by DFG projects SFB 732 and WordGraph. We also thank the anonymous reviewers for their comments.

References

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic extraction of arabic multiword expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions*, pages 19–27, Beijing, China. Coling 2010 Organizing Committee.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics.

Helena Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2009. Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the 2009 Workshop on Multiword Expressions*, pages 1–8, Singapore. Association for Computational Linguistics.

Yaacov Choueka. 1988. Looking for needles in a haystack. In *Proceedings of RIAO88*, pages 609–623.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the 2007 on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Mona Diab and Pravin Bhutada. 2009. Verb noun construction mwe token classification. In *Proceedings of the 2009 Workshop on Multiword Expressions*, pages 17–22, Singapore. Association for Computational Linguistics.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 188–195. Association for Computational Linguistics.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the 2006 Workshop on Multiword Expressions*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Coling 2010: Posters*, pages 683–691, Beijing, China. Coling 2010 Organizing Committee.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. 2010. Developing a test collection for the evaluation of integrated search. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 627–630.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Donald Metzler and W. Bruce Croft. 2004. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40(5):735–750.

B.V. Moirón and Jörg Tiedemann. 2006. Identifying Idiomatic Expressions Using Automatic Word-Alignment. In *Multi-Word-Expressions in a Multilingual Context*, page 33.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):138–158.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh, Pennsylvania, USA. Association for Computational Linguistics.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

ChengXiang Zhai and John D. Lafferty. 2002. Two-stage language models for information retrieval. In *SIGIR*, pages 49–56. ACM.