Predicting Cognitively Salient Modifiers of the Constitutive Parts of Concepts

Gerhard Kremer and **Marco Baroni**
CIMeC, University of Trento, Italy
(gerhard.kremer|marco.baroni)@unitn.it

Abstract

When subjects describe concepts in terms of their characteristic properties, they often produce *composite* properties, e. g., rabbits are said to have *long* ears, not just ears. We present a set of simple methods to extract the modifiers of composite properties (in particular: parts) from corpora. We achieve our best performance by combining evidence about the association between the modifier and the part both within the context of the target concept and independently of it. We show that this performance is relatively stable across languages (Italian and German) and for production vs. perception of properties.

1 Introduction

Subject-generated concept descriptions in terms of properties of different kinds (category: *rabbits* are *mammals*, parts: they have *long ears*, behaviour: they *jump*, ...) are widely used in cognitive science as proxies to feature-based representations of concepts in the mind (Garrard et al., 2001; McRae et al., 2005; Vinson and Vigliocco, 2008). These *feature norms* (as collections of subject-elicited properties are called in the relevant literature) are used in simulations of cognitive tasks and experimental design. Moreover, vector spaces that have subject-generated properties as dimensions have been shown to be a good complement or alternative to traditional semantic models based on corpus collocates (Andrews et al., 2009; Baroni et al., 2010).

Since the concept–property pairs in feature norms resemble the tuples that relation extraction algorithms extract from corpora (Hearst, 1992; Pantel and Pennacchiotti, 2006), recent research has attempted to extract feature-norm-like concept descriptions from corpora (Almuhareb, 2006; Baroni et al., 2010; Shaoul and Westbury, 2008). From a practical point of view, the success of this enterprise would mean being able to produce much larger norms without the need to resort to expensive and time-consuming elicitation experiments, leading to wider cognitive simulations and possibly better vector space models of semantics. From a theoretical point of view, a corpus-based system that produces human-like concept descriptions might provide cues of how humans themselves come up with such descriptions.

However, the corpus-based models proposed for this task up to this point overlook the fact that subjects very often produce *composite* properties: Subjects state that rabbits have *long* ears, not just ears; cars have *four* wheels; a calf is a *baby* cow, etc. Composite properties are not multi-word expressions in the usual sense. There is nothing special or idiomatic about *long ears*. It is just that we find it to be a remarkable fact about rabbits, worth stating in their description, that their ears are long. In the norms described in section 3, around one third of the part descriptions are composite. Note that while our focus is on feature norms, a similar point about the importance of composite properties could be made for other knowledge repositories of importance to computational linguistics, such as WordNet (Fellbaum, 1998) and ConceptNet (Liu and Singh, 2004), approximately 68,000 (36%) of the entries and 1,300 (32%) of the part entries being composites, respectively.

In this paper, we tackle the problem of generating composite properties from corpus data by simplifying it in various ways. First, we focus on *part* properties only, because they are commonly encountered in feature norms, and because they are are commonly composite (cf. section 3). Second, we assume that an early step in the process of property extraction has already generated a list of simple parts, perhaps using an existing whole–part relation extraction algorithm (Girju et al., 2006). Finally, we focus on composite parts

Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010, pages 54–62,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

with an *adjective–noun* structure – together with *numeral–noun* cases, these constitute the near totality of composite parts in the norms described in section 3. Having thus delimited the scope of our exploration, we will adopt the following terminology: *concept* for the target nominal concept (*rabbit*), *part* for the (nominal) part property (*ear*) and *modifier* for the adjective that makes the part composite (*long*).

We present simple methods that, given a list of concept–part pairs and a POS-tagged and lemmatised corpus, rank and extract candidate modifiers for the parts when predicated of the concepts. We exploit the co-occurrence patterns of the part with the modifier both near the concept and in other contexts (both kinds of co-occurrences turn out to be helpful). We first test our methods on German feature norms, and then we show that they generalise well by applying them to similar data in Italian, and to the same set of German concept–part pairs when evaluated by asking new subjects to rate the top ranked modifiers generated by the ranking methods. This also leads to a more general discussion of differences between modifiers produced by subjects in the elicitation experiment and those that are rated acceptable in perception, and the significance of this for corpus-based property generation.

The paper is structured as follows. After shortly reviewing some related work in section 2, in section 3, we describe our feature norms focusing in particular on composite properties. In section 4, we describe our methods to harvest modifiers from a corpus and report the extraction experiments, whereas section 5 concludes by discussing directions for further work.

2 Related Work

We are not aware of other attempts to extract concept-dependent modifiers of properties. We review instead related work in feature norm collection and prediction, and mention some relevant literature on the extraction of significant co-occurrences from corpora.

Feature-based concept description norms have been collected in psychology for decades. Among the more recent publicly available norms of this sort, there are those collected by Garrard et al. (2001), Vinson and Vigliocco (2008) and McRae et al. (2005). The latter was the main methodological inspiration for the bilingual norms we rely on (see section 3 below). The norms of McRae and colleagues include descriptions of 541 concrete concepts corresponding to English nouns. The 725 subjects that rated these concepts had to list their features on a paper questionnaire. The produced features were then normalised and classified into categories such as *part* and *function* by the experimenters. The published norms include, among other kinds of information, the frequency of production of each feature for a concept by the subjects.

Almuhareb (2006) was the first to attempt to reproduce subject-generated features with text mining techniques. He computed precision and recall measures of various pattern-based feature extraction methods using Vinson and Vigliocco's norms for 35 concepts as a gold standard. The best precision was around 16% at about 11% recall; maximum recall was around 60% with less than 2% precision, confirming how difficult the task is. Importantly for our purposes, Almuhareb removed the modifier from composite features before running the experiments (*1 wheel* converted to *wheel*), thus eschewing the main characteristic of subject-generated concept descriptions that we tackle here. Shaoul and Westbury (2008) and Baroni et al. (2010) used corpus-based semantic space models to predict the top 10 features of 44 concepts from the McRae norms. The best model (Baroni et al.'s Strudel) guesses on average 24% of the human-produced features, again confirming the difficulty of the task. And, again, the test set was pre-processed to remove modifiers of composite features, thus sidestepping the problem we want to deal with. It is worth remarking that, by removing modifiers, previous authors are making the task easier in terms of feature extraction procedure (because the algorithms only need to look for single words), but they also create artificial "salient" features that, once the modifier has been stripped of, are not that salient anymore (what distinguishes a monocycle from a tricycle is that one has 1 wheel, the other 3, not simply having wheels). It is conceivable that a method to assign sensible modifiers to features might actually improve the overall quality of feature extraction algorithms.

Following a very long tradition in computational linguistics (Church and Hanks, 1990), we use co-occurrence statistics for words in certain contexts to hypothesise a meaningful connection between the words. In this respect, what we propose is not different from common methods to extract and rank collocations, multi-word expressions or semantically related terms (Evert, 2008). From a technical point of view, the innovative aspect of our task is that we do not just look for co-occurrences between two items, but for co-occurrences in the context of a third element, i. e., we are interested in modifier–part pairs that are related when predicated of a certain concept. The method we apply to the extraction of modifier–part pairs when they co-occur with the target concept in a large window is similar to the idea of looking for partially untethered contextual patterns proposed by Garera and Yarowsky (2009), that extract name–pattern–property tuples where the pattern and the property must be adjacent, but the target name is only required to occur in the same sentence.

3 Composite Parts in Feature Norms

Our empirical starting point are the feature norms collected in parallel from 73 German and 69 Italian subjects by Kremer et al. (2008), following a methodology similar to that of McRae et al. (2005). The norms pertain to 50 concrete concepts from 10 classes such as mammals (e. g., *dog*), manipulable tools (e. g., *comb*), etc. The concept–part pairs in these norms served on the one hand as input to our algorithm – on the other hand, its output (the set of selected modifiers from the corpus) could be evaluated against those modifiers that were produced by the subjects. Furthermore, the bilingual nature of the norms allows us to tune our algorithm on one language (German), and evaluate its performance on the other (Italian), to assess its cross-lingual generalisation capability.

To confirm that speakers actually frequently produce properties composed of part and modifier, observe that in the German data (10,010 descriptive phrases in total), of the 1,667 parts produced, 625 (more than one third) were composite parts, and 404 were composed of an adjective and a noun, the target of this research work. Looking at the distinct parts that were elicited, 92 were always produced with a modifier, 280 only without modifier, and 122 both with and without modifier. That is, for about 43% of the parts at least some speakers used a composite expression of adjective and noun. This high proportion motivates our work and is not surprising, given that, for describing a specific concept, one will tend to come up with whatever makes this concept special and distinguishes it from other concepts – which (considering parts) sometimes is the part itself (*elephant: trunk*) and sometimes something special about the shape, colour, size, or other attributes of the part (*elephant: big ears*).

The data set for modifier extraction and subsequent method evaluation comprises all the concept–modifier–part triples (e. g., *onion: brown peel*) produced by at least one subject, taken from the German and the Italian norms. The German (Italian) speakers described 41 (30) different concepts by using at least one out of 80 (45) different parts in combination with one out of 62 (50) different modifiers, totalling to 229 (127) differently combined triples.

4 Experiments

This section describes the approach we explored for ranking and extracting modifiers of composite parts and evaluates the performance of 6 different extraction methods in terms of the production norms. Acceptance rate data from a follow-up judgement experiment complete the evaluation.

4.1 Ranked Modifier Lists

Based on the idea that the co-occurrence of words in a text corpus reflects to some extent how strong these words are associated in speakers' minds (Spence and Owens, 1990), our extraction approach works on the lemmatised and POS-tagged German WaCky[1] web corpus of about 1.2 billion tokens.

Modifier–Part Frequencies

Using the CQP[2] tool, corpus frequencies were collected for all co-occurrences of adjectives with those part nouns that were produced in the experiment described above. A possible gap of up to 3 tokens between the pair of adjective and noun allowed to extract also adjectives that are not directly adjacent to the nouns in the corpus (but in a sequence of adjectives, for example). For each part noun, the 5 most frequent adjective modifiers from the ranked modifier–part list were selected under the assumption that the preferred usage of these modifiers with the specific part indicates the most common attributes which that part typically has.

[1] See the WaCky project at `http://wacky.sslmit.unibo.it`

[2] Corpus Query Processor (part of the IMS Open Corpus Workbench, see `http://cwb.sourceforge.net`)

Log-Likelihood Values of Frequencies

An attempt to improve the performance of the first method is to calculate[3] the log-likelihood association value for each modifier–part pair instead of keeping the raw co-occurrence frequency, and select the 5 highest ranked modifiers for each part from this list. Log-likelihood weighting should account for typical modifiers which have a low frequency but do generally not occur often in the corpus, and with not many other parts – their log-likelihood value will be higher, and so will be their rank (e. g., *two-sided blade* in contrast to *long blade*).

Modifier–Part Frequencies in Concept Context

However, both of these methods do not necessarily yield generally atypical modifiers that are however typical of a part when it is attributed to a specific concept. For example, birds' beaks are typically brown, orange or yellow, but aiming to extract modifiers for a crow's beak, *black* would be one of the desired modifiers – which does not appear at a high frequency rank as a generic beak modifier. The methods described so far did not take the concept into account when generating the modifier–part pairs, i. e., for all concepts with a specific part the same set of modifiers would be extracted.

To address this issue, a second frequency rank list was prepared in the same manner – with the only difference that the part noun had to appear within the context of the concept noun. That way, also modifiers for specific concepts' parts that deviate from the most typical part modifiers appear at a high rank. However, these data are sparser, which is why we used a wide context of 40 sentences (20 sentences before and after the part) within which the concept had to occur (i. e., a paragraph-like context size in which the topic, presumably, comprises the concept). We refer to ranked lists of modifier–part pairs that do not take the target concept into account as contextless lists, and to lists within the span of a context as in-context lists.

Due to the already mentioned data sparseness problem, not all modifiers used for a part noun in the production norms could be extracted with the latter method, as some of the obvious modifiers for specific parts are just not written about. For these, there is a higher chance that they appear, if at all, in the contextless rank list. For example, *thin bristles* does not appear in the context of *broom*. In the in-context list, 33% of the 229 triples extracted from the German norms were not found (in the contextless list, only 9% modifier–part pairs are missing). Additionally, particular concepts, parts, or concept–part pairs (within the 40 sentence span) might be missing from the corpus, as well. From the German norms collection, all concepts appeared in the corpus, but one part (a noun–noun compound), and 6 concept–part pairs (rare, colloquial part nouns) were missing. In the evaluation to follow, all the modifiers pertaining to these missing data from the corpus will be counted as positives not found by the algorithm.

	contextless		concept context	
rank	freq	modifier	freq	modifier
1	507	thick	16	thick
2	209	dense	14	white
3	204	soft	11	small
4	185	black	11	soft
5	175	long	9	dense

Table 1: Top 5 modifiers from frequency rank lists for part *fur* and concept *bear*

The example excerpt in table 1 shows modifiers that were selected for *bear* and *fur*, using the two frequency rank lists described above. Although in this example most of the modifiers (thick, dense, soft) are found in both lists, two arguably reasonable modifiers are just in the contextless set (black, long), and one only in the in-context set (white). A disadvantage of selecting modifiers from the in-context rank list is that many modifiers have the same low frequency, but they should nevertheless have differing ranks. In such cases, we assigned ranks according to alphabetic order of modifiers.

Summed Log-Rescaled Frequencies

Next, to improve performance and profit from both information sources the above methods provide, the in-context and contextless rank lists were combined. In one variant, the scaled frequencies for the concept–modifier–part triples appearing in both lists were added. Scaling was necessary because the frequencies in the contextless list are in general much higher than in the in-context list. Furthermore, to account for the fact that at high ranks the difference in frequency between subsequent ranks is much higher than at lower ranks, scaling was done by using the logarithmic values of the fre-

[3] Using the UCS toolkit, described at `http://www.collocations.de/software.html#UCS`

quencies: For each concept–modifier–part triple, its logarithmic frequency value was divided by the logarithmic value of the maximum corpus frequency of all parts in the corpus (in the contextless list) or of all concept–part pairs co-occurring within 40 sentences (in the case of the in-context list).

Productwise Combination of Frequencies

As an alternative back-off approach, the raw frequencies were combined productwise into a new list (for those modifier–part pairs missing in the in-context list, the frequency of the pair in the contextless list was taken alone, instead of multiplying it by zero; i. e., the in-context term was $\max(\text{freq}, 1)$). This achieves a sort of "intersective" effect, where modifiers that are both commonly attributed to the part and predicated of it in the context of the target concept are boosted up in the list, according to the intuition that a good modifier should be both plausible for the part in general, and typical for the concept at hand.

Cosine-Based Re-Ranking

An attempt to further improve performance is based on the idea that parts are described by some specific types of attributes. For example, a *leaf* would be characterised by its shape or consistency (e. g., *long*, *stiff*), whereas for *fur* rather colour should be considered (e. g., *white, brown*). If we are able to cluster modifiers for their attribute type and find out which attribute types are in particular important for a specific part, those could get a preference in the rank list and be moved towards the top. To approach this in a simple way, a re-ranking method is used which is supposed to cluster and choose the right cluster of modifiers implicitly: The modifiers in the (productwise-) combined list were tested for their similarity by looking if they co-occur with the same relative frequency with the same set of nouns. In case of high similarity (in this respect) of a modifier to a single other modifier, or if the modifier was similar to a lot of modifiers, it should be re-ranked to a higher position. In more detail, a vector was created for each modifier, denoting its co-occurrence frequencies with each noun in the corpus within a window of 4 tokens (on the left side of the noun). Random indexing helped to reduce the vector dimensionality from 27,345 to 3,000 elements (Sahlgren, 2005). These vectors served for calculating the cosine distances between modifiers. Then, for each of the top 200 modifiers in the combined frequency rank list (covering 84% of the triples from the German norms), the cosine distance was calculated to each of the top 100 modifiers in the contextless rank list. A constant of 1 was added to each of the computed cosines, thus obtaining a quantity between 1 and 2. The original combined frequency value was multiplied by this quantity (thus leaving it unchanged when the original cosine was 0, increasing it otherwise). From the re-ranked list resulting from this operation, we selected, again, the top 5 modifiers of each concept–part pair. For example, suppose that *black* is among the modifiers of a *crow*'s *beak* in the combined list. We compute the cosine similarity of *black* with the top 100 modifiers of *beak* (in any context), and, for each of these cosines, we multiply the original combined value of *black* by $\text{cosine}{+}1$. Since the colour is a common attribute of beaks, the presence of modifiers like *yellow* and *brown*, high on the contextless *beak* list, helps re-ranking *black* high in the *crow*-specific *beak* list. We hope that this method helps out concept-specific *values* (e. g., *black* for *crow*) of *attributes* that are in general typical of a part (*colour* for *beak*).

4.2 Performance on Composite Parts From the Production Norms

The feature norms data represented the gold standard for the evaluation of all sets of modifiers chosen by each of the described methods for the given concept–part pairs. Note that, even if a modifier–part pair was produced only once in the feature production norms, the corresponding concept–modifier–part triple was included in the gold standard – which contains 41 different concepts, 80 different parts, and 62 different modifiers, totalling to 229 concept–modifier–part triples. As in the German corpus there are 154,935 adjective–part-noun pairs, the random baseline (random guessing) for finding these 229 pairs is approaching 0 (similarly for Italian and the judgement dataset).

Figure 1 displays the performance of the methods on German in the form of a recall–precision graph. For each rank (1–5), overall recall and interpolated precision values are given for all modifier–part pairs up to this rank – note that precision at 1% recall is overrated as it is based on an arbitrary fraction of rank 1 pairs. As expected, extracting modifiers of parts within a concept context (the in-context list) achieves low recall. In contrast, modifiers that were extracted by querying the corpus for parts without considering the concept context have

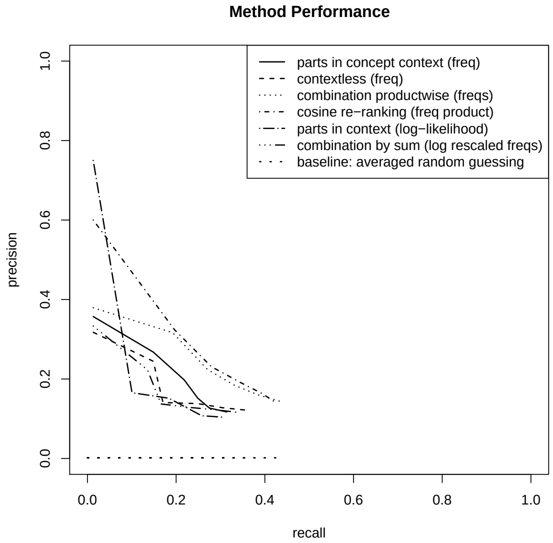

Figure 1: Evaluation on German norms

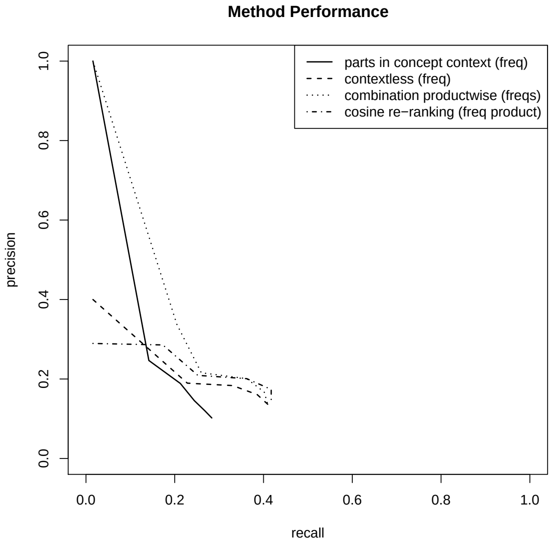

Figure 2: Evaluation on Italian norms

a higher recall. But this method has a lower precision in general. The performance for the method combining frequencies productwise and for the one that re-ranks this combined list via cosine-based smoothing are substantially better. Not only the precision is much higher at all recall levels, but also their maximum recall values are higher than those of the contextless lists, i. e., it was worth combining the complementing information in the two lists. However, the performance of the cosine-based re-ranked list compared to the productwise-combined list is not considerably higher, as we might have hoped. The remaining two alternative methods performed much worse: the one using log-likelihood values as ranking criterion had in general a low precision and a low recall, and the method combining the in-context and the contextless rank list by summing up the rescaled logarithmic frequency values performs as bad as the contextless rank list. Nevertheless, note that all methods perform distinctively well above the baseline.

Qualitatively analysing the data collected with the described methods did not give definite clues about why some performed not as good as expected. As a comprehensible example, the modifier *short* for *legs* is at rank 5 in the contextless list, but because of the frequent co-occurrence with *monkey* it rises to rank 2 in the productwise combination of these lists, and even to rank 1 in the cosine-based re-ranked list. An understandable bad performing example is the modifier *yellow* for the *eyes* of an *owl*: Although it appears in the in-context list at rank 2, it is a quite infrequent modifier for *eyes* in general (i. e., low in the contextless list), and thus it is not contained in the top 5 modifiers in the productwise combined rank list. On the other hand, it is not perfectly clear to us why, e. g., *flat* for the *roof* of a *skyscraper*, which is at rank 5 in the contextless list and at rank 6 in the combined list, is lowered to rank 9 in the cosine-based re-ranked list (in the in-context list, it does not appear at all). For all methods, collected modifiers include such of undesired attributes not describing the part, but other, rather situational aspects, e. g., *own*, *left*, *new*, *protecting*, and *famous*. Furthermore, we observed that some modifiers are reasonable for the respective concept–part pair, but they are counted as false because they did not occur in the production experiment (that we took as the evaluation basis), e. g., for the *blade* of a *sword*, not only *large* is acceptable, but also *long* and *wide*, essentially making the same assertion about the size of the *blade*. This issue is addressed further below by creating a new evaluation standard based on plausibility judgements.

To evaluate the cross-lingual performance of the extraction approach, the Italian norms were explored similarly to the German norms for composite parts. The gold standard here comprised 127 triples (from combinations of 30 different concepts, 45 parts, and 50 different modifiers). The same methods described above were used to extract modifiers from the Italian WaCky web corpus (more than 1.5 billion tokens), with one difference regarding the query for adjectives near nouns: As in the Italian language adjectives in a noun phrase can be used both before and after the noun (with differences in their meaning), and given that most of them were produced after the noun, we collected all adjectives occurring up to 2 words from the left of the noun and up to 4 words to the right.

Figure 2 shows the performance curves of the methods for the Italian data. In this evaluation, the method using log-likelihood values and the method combining lists via addition of logarithmic rescaled frequencies are omitted as their performance was not promising at all in the German data, and they are conceptually similar to the contextless and productwise-combination approaches, respectively. Like in German, the in-context method yields a low recall, in contrast to the method not considering the presence of concepts in context. Again, cosine-based re-ranking performs very similarly to the method using the productwise-combined list. For the performance on the Italian data, their difference from the simple frequency rank lists is not as large as it is for the German data, but it is clearly visible, especially at higher recall values.

Summarising, our comparison of various corpus-based ranking methods to the feature production norms, both in German and Italian, suggests that composite parts produced by subjects are best mined in corpora by making use of both general information about typical modifiers of the parts (the contextless rank) and more specific information about modifiers that co-occur with the part near the target concept. Moreover, it is better to combine the two information sources productwise, which suggests an intersecting effect (the most likely modifiers are both well-attested out of context and seen near the target concept). For both languages, there is no strong evidence that re-ranking by cosine similarity (a method that should favour modifiers that are values of common attributes of a part) is improving on the plain combination method (although re-ranking is not hurting, either).

By looking at the overall performance, the results are somewhat underwhelming, with precision around 20% at around 30% recall for the best models in both languages. A natural question at this point is whether the modifiers ranked at the top by the best methods and treated as false positives because they are not in the norms are nevertheless sensible modifiers for the parts, or whether they are truly noise. In order to explore this issue we turn now to our next experiment.

4.3 Performance Evaluation Based on Plausibility Judgements

The purpose of this judgement experiment was to see which concept–modifier–part triples the majority of participants would rate as acceptable. It allows us to investigate two topics: (i) the comparison of what people produce and what they perceive as being a prominent modifier for a concept–part pair (our algorithm might actually provide good candidates which were just not produced, as we just said) and (ii) a re-evaluation of the cosine-based re-ranking method (it could be in fact better than we thought because we only evaluated what was produced, but did not have a definite plausibility rating of the candidates missing in the norms).

The tested set contained the triples yielded by our two best performing methods (productwise combination and cosine-based re-ranking), which were applied to the German feature norms (692 triples, comprising 41 concepts and 71 parts). From this set, a set of triples was chosen randomly for each of the 46 participants (recruited by e-mail among acquaintances of the first author). The triples were presented to participants embedded into a natural-sounding sentence of the form “The [part] of a [concept] is [modifier]”. Each participant rated 333 sentences, presented on separate lines of a text file (this set of sentences presented comprised additional triples which were intended for other purposes – for the current evaluation, we used a subset of 110 of these from each participant, on the average). Participants were instructed to read the sentences as general statements about a concept’s part and mark them by typing a letter (“w” for wonderful and “d” for dubious – to facilitate one-handed typing and easy memorisation) at the beginning of the line, if they thought it plausible/unlikely that someone used the sentence to explain an aspect of the relevant part. In total, 5,525 judgements were collected; each sentence in the set was judged on the average by 8 persons.

The performance evaluation is based on the acceptance rate of the participants: Modifiers accepted by at least 75% of the raters are considered plausible. Figure 3 shows the recall–precision graph for the methods tested on the concept–part pairs from the German norms. From the 692 triples judged, around 13% were accepted by the majority of speakers. The precision rate is comparable with the evaluation on the basis of the modifiers produced by participants (highest recall is 1, of course,

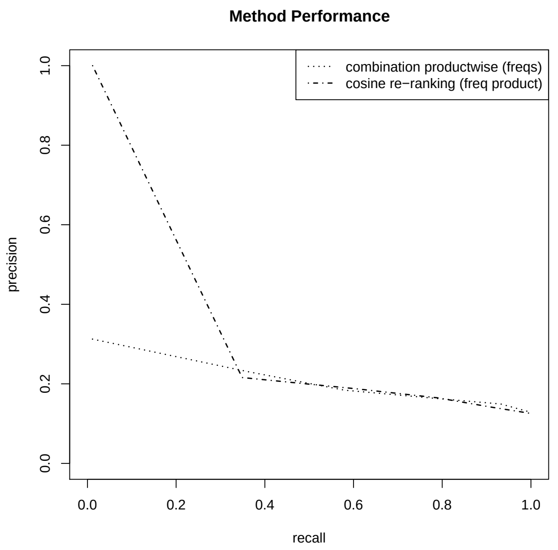

Figure 3: Evaluation on judgements (German)

because all modifiers to be judged were exclusively from the data set selected by our methods).

Again, the performance of the cosine-based re-ranking method is similar to the performance of the productwise-combination method. For a more exact evaluation of the difference between these two, a last test was conducted: Instead of measuring the performance in the form of counts of modifiers that were accepted by the majority of participants, we used the acceptance rates of all modifiers: The acceptance rates of all judged triples were summed up if they contained the same concept–part pair. This means that each concept–part pair received a score reflecting the overall acceptance of the set of modifiers for that pair (e. g., for *bear: fur*, all acceptance rates for *bear: brown fur*, *bear: soft fur*, . . . were summed up). Then, the score of each concept–part pair in the productwise-combined list was compared against the score of the same pair for the cosine-based re-ranking method, using a pairwise t-test (this procedure is sound because the modifiers per pair are the same for the two methods). The test showed a significant difference (p = 0.008), but in favour of the productwise-combination method (score means were slightly higher). That is, cosine-based re-ranking in the current form brings no advantage over the simpler productwise combination of the frequency lists.

Finally, turning to the qualitative comparison of production and perception, there was a relatively small overlap of triples (46) contrasting with modifiers only produced but not accepted (53), and modifiers accepted but not produced (42). Intuitively, we would have expected that what was produced will be also accepted by the majority of people. Possibly, some participants in the judgement experiment found a few of the triples produced questionable (*goose: long beak*) – such triples were in our gold standard because we deliberately did not want to exclude composite parts even if produced by only one speaker – whereas participants producing parts for given concepts probably just did not think of specific parts or modifiers (e. g., *aeroplane: small windows* and *bear: dense fur*). The important fact regarding this difference is, however, that our method captures both kinds of modifiers.

5 Discussion

We presented several corpus-based methods that provide a set of adjective modifiers for each concrete concept–part pair, to be compared to those modifiers that are salient to human subjects. The general approach was to generate ranked lists, and select the 5 candidates at the top of the ranks.

The best of our methods works on the simple (productwise-) combination of frequency information of co-occurring adjective–noun pairs with and without considering a wide “concept context” in which the part noun has to occur. This method performed better than the one based on co-occurrence frequency not in concept context (generic modifiers, not appropriate for every concept) and the one based on co-occurrence frequencies in concept context, only (low recall because of sparse data).

We evaluated the methods on feature production norms and on plausibility judgements of generated concept–modifier–part triples to compare production and perception of modifiers. The performance was similar in precision – although the qualitative analysis showed that modifiers produced and modifiers perceived did not have a large overlap. This means our algorithm is capable of collecting both with the same performance.

After tuning the algorithm on German norms, we evaluated its generalisation capability to a different language (Italian). Performance was similar. Less satisfying at first glance is the precision value of just around 20% at the maximum recall level (however, when compared to the baseline of below 1% precision, this is an essentially better value) – as well as the fact that our implementation of the intuitive idea to re-rank modifiers that are similar (and should instantiate the same attribute) did not have a performance advantage. This is subject to further work. Moreover, using a machine-learning method (building a binary classifier) could be tried. Another idea was to crawl the web and select concept-specific text passages to build a specialised corpus. Possibly, we could draw then from a richer information source. A rough attempt to do this did not seem to yield promising results.

So far, we included only adjectives as permissible modifiers. A future extension could be also aiming for numerals (e. g., *four wheels*). Then, for the simulation of human-like behaviour we imagine as part of the possible future work to enable the algorithm to decide if a part noun should be paired with a modifier, at all – or if the part itself is sufficient to describe a concept (*big ears* vs. *trunk*).

Regarding the evaluation, a more exact performance measure would probably be achieved by either having more participants producing concept descriptions and then only selecting those modifiers for the gold standard that were produced by a majority – or letting participants in a judgement experiment also judge modifiers that were produced, to filter out the unlikely ones.

A next step in the project will be extracting salient parts for concepts (which we assumed to have done already for the purpose of this paper), possibly by integrating the information we already collected by extracting modifiers. In the end, we would like to come up with an adaptable method that extracts not only parts but also other types of relations (e. g., category, behaviour, function, etc.), which have been already addressed in related works, though. The issue we presented in this paper, however, is new and, we think, worth exploring.

References

Abdulrahman Almuhareb. 2006. *Attributes in Lexical Acquisition*. Phd thesis, University of Essex.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Kenneth Church and Peter Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin, Germany.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Nikesh Garera and David Yarowsky. 2009. Structural, transitive and latent models for biographic fact extraction. In *Proceedings of EACL*, pages 300–308, Athens, Greece.

Peter Garrard, Matthew Lambon Ralph, John Hodges, and Karalyn Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2):25–174.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*, pages 539–545, Nantes, France.

Gerhard Kremer, Andrea Abel, and Marco Baroni. 2008. Cognitively salient relations for multilingual lexicography. In *Proceedings of the COGALEX Workshop at COLING08*, pages 94–101.

Hugo Liu and Push Singh. 2004. ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, pages 211–226.

Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of COLING-ACL*, pages 113–120, Sydney, Australia.

Magnus Sahlgren. 2005. An introduction to random indexing. `http://www.sics.se/~mange/papers/RI_intro.pdf`.

Cyrus Shaoul and Chris Westbury. 2008. Performance of HAL-like word space models on semantic clustering. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 42–46, Hamburg, Germany.

Donald Spence and Kimberly Owens. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5):317–330.

David Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.