Supervised Learning of German Qualia Relations

Yannick Versley
SFB 833
Universität Tübingen
versley@sfs.uni-tuebingen.de

Abstract

In the last decade, substantial progress has been made in the induction of semantic relations from raw text, especially of hypernymy and meronymy in the English language and in the classification of noun-noun relations in compounds or other contexts. We investigate the question of learning qualia-like semantic relations that cross part-of-speech boundaries for German, by first introducing a hand-tagged dataset of associated noun-verb pairs for this task, and then provide classification results using a general framework for supervised classification of lexical relations.

1 Introduction

Ever since the introduction of wordnets (Miller and Fellbaum, 1991) or more generally machine-readable dictionaries containing semantic relations, researchers have investigated ways to learn such examples automatically from large text corpora, or generalize them from existing instances. Substantial research exists on the learning of hyperonymy relations (Hearst, 1992; Snow et al., 2005; Tjong Kim Sang and Hofmann, 2009), meronymy relations (Hearst, 1998; Berland and Charniak, 1999; Girju et al., 2003) and selectional preferences (Erk et al., 2010; Bergsma et al., 2008; Ó Séaghdha, 2010).

Both lexicographic research (Chaffin and Herrmann, 1987; Morris and Hirst, 2004) and research in cognitive psychology (Vigliocco et al., 2004; McRae et al., 2005), argue that it is important to consider relations beyond the classical inventory of hyperonymy and meronymy relations; furthermore psychological research on priming (Hare et al., 2009) suggests different processing for different relations, which would entail that cognitively plausible modeling of human language should model these relations explicitly rather than simply recording untyped associations between concepts (as in the 'evocation' relation proposed for WordNet by Boyd-Graber et al., 2006).

One set of suggestions for an extended inventory of relations can be found in the telic and agentive qualia relations of Pustejovsky (1991) which have been shown to be useful in recognizing discourse relations (Wellner et al., 2006), or metonymy/coercion phenomena (Verspoor, 1997; Rüd and Zarcone, 2011), and have the property of linking different parts-of-speech groups, unlike meronymy and hyperonymy/troponymy.

The work we present in this paper consists of a dataset of noun-verb associations for German concrete nouns, which we present in more detail in section 3, and a state-of-the-art approach to the supervised classification of such cross-part-of-speech relations using informative features from large collections of unannotated text, which we present in section 4. Experimental results are discussed in section 6.

2 Related Work

Most of earlier work on discovering novel instances of semantic relations was based on surface pattern matching, as presented by Hearst (1998). In the domain of finding qualia relations, Cimiano and Wenderoth (2005) propose patterns such as "*...purpose*

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 12–23,
Jeju, Republic of Korea, 12 July 2012.

of X is …" or "*…X is used to …*", whereas they argue that agentive qualia are best chosen from a small, fixed inventory of verbs (e.g., *make*, *bake*, *create …*). Katrenko and Adriaans (2008a) additionally propose "*to Y a (new|complete) X*" and "*a (new|complete) X has been Y'd*" as patterns for agentive qualia.

Some of the more recent work starts out from matches extracted by means of such a pattern, but use supervised training data to learn semantic constraints that improve the precision by filtering the extracted examples. Berland and Charniak (1999) use some handcrafted rules to exclude abstract objects from the part-of relations they extract from a corpus, and additionally rank pattern extractions by collocation strength. Girju et al. (2003) propose an iterative refinement scheme based on taxonomic information from WordNet: In this learning approach, general constraints using top-level semantic classes (entity, abstraction, causal-agent) are passed to a decision tree learner and iteratively refined until the semantic constraints induced from the classes are no longer ambiguous.

Katrenko and Adriaans (2008b, 2010) present approaches to learn semantic constraints for the use in recognizing semantic relations between word tokens (SemEval 2007 shared task, see Girju et al., 2009), either in a graph-based generalization of Girju's iterative refinement approach that is able to handle sense ambiguities more gracefully, or by clustering pairs of words by the joint similarity of both relation arguments.

A complementary aspect is to improving recall beyond the possibilities of a few hand-selected patterns. Following Hearst (1998), Girju et al. (2003) show that it is possible to find usable patterns by exploiting known positive examples and looking for co-occurrences of these relation arguments in a corpus. However, these patterns usually have low precision and/or very limited recall, meaning that a more elaborate approach (such as Girju et al.'s induction of semantic constraints) is needed to make the best use of them.

Yamada and Baldwin (2004) propose to use a combination of templates typical of telic and agentive qualia relations (X is worth Ying, X deserves Ying, a well-Yed X) and a statistical ranking combining association and a classifier learned on positive and negative examples for that role. They find that the combination of association statistic and classification worked somewhat better than the templates alone.

One approach targeted at exploiting a greater number of patterns for hyperonymy relations can be found in the work of Snow et al. (2005): they extract patterns consisting of the shortest path in the dependency graph plus an optional satellite and use the set of all found paths as features in a linear classifier. The resulting classifier for hyperonymy relations outperforms single patterns both in terms of precision and in terms of recall; a further improvement can be achieved if the frequency of pattern instances is binned instead of just occurrence or non-occurrence being recorded.

Tjong Kim Sang and Hofmann (2009) investigate the question whether it is necessary to use syntactic (rather than surface) patterns for the hyperonym classification approach of Snow et al. They compare a method of extracting features based on syntax as in Snow et al.'s approach with a surface-based alternative where the string between two words, plus optionally one word to the left or right side of the word, is extracted. Tjong Kim Sang and Hofmann argue that the benefit of the parser (additional recall due to the better generalization capability of the syntactic patterns) is mostly negated by parsing errors: In some informative contexts that the system based on POS patterns is able to find without problems, parsing errors lead to a parse tree that does not exhibit the intended (dependency path) pattern.

Several researchers have applied such pattern classification approaches to a larger set of relations, and have demonstrated that extracting a pattern distribution between occurrences and performing supervised classification based on this distribution is a promising solution for semantic relations that go beyond hyperonymy.

Ó Séaghdha and Copestake (2007) use a supervised classification approach for noun-noun compounds combining context features for each of the single words with features characterizing the joint occurrences of the two nouns that are part of the target compound. In their experiments, they found that linear classification using informative (bag-of-words and bag-of-triples) features in conjunction with features aimed at the similarity of each word of the target pair yields good results. In particular, the results of using a linear classifier with informative corpus-based features that are quite close to those that can be achieved using a (more accurate, but computationally quite expensive) string kernel or those that Ó Séaghdha (2007) achieves using taxonomic information from WordNet.

Turney (2008) presents a general approach for classifying word pairs into semantic relations by extracting the strings occurring between the two words of a pair (up to three words in-between, up to one word on either side) and using a frequency-based selection process to select sub-patterns where words from the extracted context pattern may have been replaced by a wildcard. Using standard machine learning tools (a support vector machine with radial base function kernel), he is able to reach results that are close to those possible with previous more specialized approaches.

Similarly, Herdağdelen and Baroni (2009) tackle a variety of problems in semantic relation classification using a unified approach where frequent unigrams and bigrams are extracted from co-occurrence contexts of the target word pair (in addition to features extracted from general occurrence contexts of each word). Herdağdelen and Baroni's approach uses a linear SVM (which is faster and better-suited to large data sets in general than either kernelized support vector machines or nearest-neighbour approaches) yet is able to reach competitive accuracy.

In contrast to approaches using generic machine learning, Ó Séaghdha and Copestake (2009) and Nakov and Kozareva (2011) model the similarities between related word pairs more explicitly in terms of distributional kernels (Ó Séaghdha and Copestake), or as a similarity metric between word pairs (Nakov and Kozareva). Such approaches allow more flexibility in the modeling of similarity and the combination of lexical and relational similarity measures, but are less well-suited for scaling up to more training data.[1]

Because of the need for sufficient training data, purely supervised approaches to learning relations in morphologically-rich languages are often limited to the classical relations found in wordnets. Tjong Kim Sang and Hofmann (2009) use a Dutch corpus and hyperonymy relations from the Dutch Cornetto wordnet and mention relatively few differences to approaches on English such as Snow et al. (2005). Kurc and Piasecki (2008) apply the semi-supervised approach of Pantel and Pennachiotti (2006) for learning hyperonymy relations, but modify the patterns used to enforce morphosyntactic agreement and accommodate a more flexible word order. Versley (2007) uses Web pattern queries for finding hyperonymy relations and mentions the fact that greater morphological richness and the smaller size of the German Web make the use of Web queries more complex than for English.

Outside the realm of hyperonymy, Regneri (2006) uses Web-based pattern search to classify verb-verb associations into the semantic classes proposed for English by Chklovski and Pantel (2004). Rüd and Zarcone (2011) perform a corpus study of patterns indicative of telic and agentive qualia relations in a German Web corpus, but perform no automatic classification.

In summary, the research of Tjong Kim Sang and Hofmann (2009) seems to indicate that at least hyperonymy relations can be found using a shallow pattern approach despite greater word order flexibility of languages such as Dutch and German. For cross-part-of-speech relations, such as telic and agentive qualia, such a question has been unaddressed as of yet, which prompted us to create a dataset that is suitable for such an investigation.

3 Material

In order to investigate general-domain Noun-Verb relations in German, we first had to create an appropriate dataset that captures a realistic notion of the relationships that humans infer in a text. Existing datasets that explore this space (most of them for English) use a variety of approaches: One approach starts from examples (such as the popular analogy dataset for English introduced by Turney and Littman, 2003); other approaches such as the data collection for the SemEval task on identifying relations between nominals (Girju et al., 2009; Hendrickx et al., 2010) start from common semantic re-

[1] Ó Séaghdha and Copestake (2009) reports training times of slightly more than one day for their most efficient method whereas a ten-fold crossvalidation run using SVM$_{\text{perf}}$ – see the presentation on p. 6 – takes under an hour, i.e., using linear classification is more efficient by a factor of about 100.

lations and use patterns to gather positive and negative examples by Web queries.

In our case, we started from noun-verb associations found in a sample of human-produced associations to concrete noun stimuli (Melinger et al., 2006); starting from the original association data, we excluded items that were produced by less than three subjects and used the part-of-speech information attached to the data to retrieve only the verb associates.

The classification scheme was motivated by existing generative lexicon research (Pustejovsky, 1991; Lenci et al., 2003), but was modeled to achieve a good fit to the associations present in the data rather than to force a good fit to any particular theory.

- *agentive* relations exist between an artifact and an event that creates or procures it (e.g. *bread-bake*)
- the *telic* relations exist between an entity and an event that is related to its purpose or (actual or intended) role:
  - *telic-artifact* holds between an artifact and its intended usage (e.g. *plane-fly*)
  - *telic-role* holds between a role (i.e., a profession, organizational position etc.) and activities related to that role (e.g. *cowboy-ride*)
  - *telic-bodypart* holds between a body part and its intended uses (e.g. *eye-see*)
- the *behaviour* group of relations hold between an entity and events that are caused by it, but are not necessarily intentional or related to a role that it fulfills:
  - *behaviour-animate* are typical activities performed by animate entities that are unrelated to the role that they fulfill for humans (e.g., *dog-bark*)
  - *behaviour-artifact* relates artifacts to (usually) unintended behaviour associated with them (e.g., *moped-rattle*)
  - *behaviour-environment* relates elements of the environment to events that go on around them (e.g., *sun-shine*)
- *location* relations hold between elements of the environment and activities typically performed in or at them (e.g., *mountain-climb*)
- *grooming* relations hold between artifacts and activities that contribute to the readiness of an artifact (or body part) for its intended use but are not directly related to it (e.g., *plant-water*, *hair-dye*)

In comparison to standard schemes such as SIMPLE (Lenci et al., 2003), we have extended the set of *telic* and *agentive* qualia from the original generative lexicon approach by supplementing it with relations that describe the affordances of objects or guides the interpretative linking of objects and events, namely *location* for affordances of elements of the environment and *grooming* for object-related actions that may not be necessary for a differently-built object with that same function, and finally *behaviour* describes events that co-occur with objects but are usually not part of a human agent's action plan.

As a refinement, we subdivided the *telic* qualia and *behaviour* relations, in particular specifying any *telic* relation with the reason a concrete object may be relevant for goal-directed processing – either by teleological interpretation of body parts, by the creation of artifacts with a specific purpose, or the establishment of roles with social conventions supporting certain types of actions.

Among the responses collected by Melinger et al. (2006), we found relatively few instances that were genuinely ambiguous (*Drachen - fliegen*, which may either be interpreted as 'kite/fly', in which case it would be a *telic-artifact* relation, or as 'dragon/fly', in which case it would be a *behaviour-animate* relation), but found that domestic animals (cows, horses, dogs) have affordances such as *horse-ride* or *dog-bark* that indicate they are conceptualized as instruments serving a particular goal (which means that the relation should be labeled as *telic-artifact* rather than as *behaviour-animate*).

In the associated word pairs, we also found relations such as *Zwiebel-schneiden* ('onion-cut') or *Handtuch-duschen* ('towel-shower') where the action is related to a thing's purpose but not identical to it (towels are used to dry yourself *after* showering, and people acquire onions to eat them *after* having cut them). Our initial annotation included a combination between the qualia-like relations presented here and an additional event-semantic relation linking the elicited event and the intended affordance of the object. However, the event relation was left out of the dataset used in the experiments to avoid data sparsity.

In our dataset with 641 items, the most frequent relations are *telic-artifact* (425 instances), *behaviour-animate* (94 instances), *telic-role* (35 instances), *telic-bodypart* (24 instances). The other relations have between 2 and 17 instances each (see table 3). The relationship data is therefore heavily skewed.

4 Classification Approach

Our classification approach is aimed at a practical toolkit for supervised classification of lexical-semantic relations, similar in spirit to the BagPack approach of Herdağdelen and Baroni (2009) but adapted for the use in morphologically-rich languages, in particular German.

In addition to the surface-based unigram and bigram features, we use features based on dependency syntax, which is more robust against variation in word order, and allows to reattach separable verb prefixes.

4.1 Preprocessing

To see why a very shallow approach may be less useful for German, let us consider a simple direct (accusative) object relation such as between *aufessen* (eat up) and *Kuchen* (cake): this relation could be realized in a variety of ways depending on clause type and constituent order, as illustrated in example (1).

(1) a. Peter isst den Kuchen auf.
Peter eats the cake up.
"Peter eats up the cake".

b. Den Kuchen hat Peter aufgegessen.
The cake$_{acc}$ *has Peter eaten-up.*
"Peter has eaten up the cake".

c. . . . dass Peter den Kuchen aufisst.
. . . that Peter the cake up-eat.
". . . that Peter eats up the cake".

In German, clause type decides whether the verb is in verb-second position (1a) or at the end of the clause (1b,1c); additionally, as in (1a), prefixes of verbs may be stranded at the end of a clause with the verb in verb-second position.

In addition to morphological analysis, hence, reattachment is necessary in such cases as (1a), and parsing is necessary to reattach prefix and verb. In cases such as (1b), word order variation also needs to be taken into account in order to recover the direct object relation, unlike in languages with less-flexible word order.

As a text collection that furnishes contexts for the words or word pairs that interest us, we use the *web-news* corpus, a collection of online news articles collected by Versley and Panchenko (2012). For the processing of this 1.7 billion word corpus, we use a pipeline that relies on deterministic dependency parsing to provide complete dependency parses at a speed that is suitable for the processing of Web-scale corpora.

The parsing model is based on MALTParser, a transition-based parser, and uses part-of-speech and morphological information as input. Morphological information is annotated using RFTagger (Schmid and Laws, 2008), a state-of-the-art morphological tagger based on decision trees and a large context window (which allows it to model morphological agreement more accurately than a normal trigram-based sequence tagger). While transition-based parsers are quite fast in general, an SVM classifier (which is used in MALTParser by default) becomes slower with increasing training set. In contrast, using the MALTParser interface to LibLinear by Cassel (2009), we were able to reach a much larger speed of 55 sentences per second (against 0.4 sentences per second for a more feature-rich SVM-based model that reaches state of the art performance).

For lemmatization, we use the syntax-based TüBa-D/Z lemmatizer (Versley et al., 2010), which uses a separate morphological analyzer and some fallback heuristics. The SMOR morphology (Schmid et al., 2004) serves to provide morphological analyses for novel words, covering inflection, derivation and composition processes. For unanalyzed novel words that are not covered by SMOR, the lemmatizer falls back to surface-based guessing heuristics. It uses morphological and syntactic information to provide more accurate lemmas; In addition to dependency structures, the morphological tags from RFTagger as well as global frequency information are used.

4.2 Classification

For classification, we use the following learning methods:

- For the **SVMperf** classifier, the set of possible labels is decomposed into binary problems using the one-vs-all scheme (for each possible label, a classifier is trained that receives the instances of this label as positive instances and the others as negative instances). SVMperf allows the training of models that either optimize (an upper bound for) the accuracy ($\mathrm{SVM}_{\mathrm{acc}}$) or the f-measure ($\mathrm{SVM}_{\mathrm{F}}$) of positive instances (Joachims, 2005).

- The **Maximum Entropy** (MaxEnt) classifier directly learns the multiclass decision. Here, we used the AMIS package by Miyao and Tsujii (2002).

All experiments are run in a ten-fold crossvalidation setup where the data is split ten portions and each portion (fold) is tagged using a classifier trained on the remaining nine folds. This setup leads to decreased variation

As noted in section 6, $\mathrm{SVM}_{\mathrm{perf}}$ using optimization for accuracy (i.e., a standard linear kernel SVM with hinge loss and a one-versus-all reduction to handle the multiclass problem) performs best on the two aggregate measures that we used (accuracy and macro-averaged F). Hence, most results we report in the later part only use the standard SVM learner.

4.3 Features

The first group of **surface-based** features uses a similar technique to Herdağdelen and Baroni (2009): given the co-occurrences of two words X and Y with at most 4 words in-between, we extract frequent unigrams and bigrams. Because we can maintain the sparsity of the resulting feature vector (see section 5), we can use a larger list of 10 000 each of the most frequent unigrams and bigrams (`w12`) alternatively to a list with only 2 000 entries each (`w12:2k`). The `lem12` feature uses the same approach, but uses lemmas instead.

A second group of features uses a **path-based** representation based on a modified version of the dependency parse (where the main verb, and not the auxiliary verb is the head of a clause and is connected to both the subject and its other arguments).

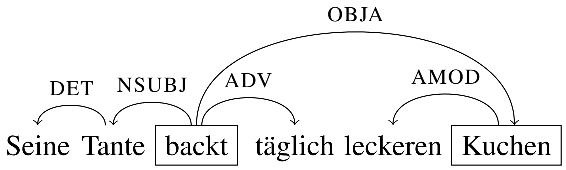
In the path-based representation, we can extract the (shortest) path between the two target words in the dependency graph. The `rel` feature records the complete path (labeled dependency edges as well as lemmas of intervening nodes) between the target words. In contrast, the `sat` feature records labeled dependency edges as well as lemmas of the dependents of one of the target words.

Because the `rel` feature yields relatively large (and therefore sparse) strings, we also decompose the dependency path in triples consisting of labeled dependency edge in the path and the two nodes adjacent to it (with the endpoints replaced by “w_1” and “w_2”, respectively) for the `triples` feature.

In order to emulate the feature extraction of Snow et al. (2005), we introduce a `relsat` feature, which pairs the path (as in the `rel` feature) with one dependent of either target word. The `relsat` feature would be able to model patterns such as “w_1 and other w_2”, where a modifier (“other”) is not part of the shortest dependency path between w_1 and w_2.

In addition, a feature based on GermaNet (Henrich and Hinrichs, 2010) uses **taxonomic** information: possible hypernyms of the noun and verb in the pair are extracted, and are used by themselves (e.g. “noun is a hyponym of ‘thing’ ”, or “verb is a hyponym of ‘communicate’ ”) and in combinations of up to two of these possible hypernym labels.

In addition to taxonomic information from GermaNet, we use **distributional similarity** features for single words. For the nouns, we use distributional features based on the co-occurrence of premodifying adjectives, which Versley and Panchenko (2012) found to work better than other grammatical-relation-based collocates (`attr1`), while we use Padó and Lapata’s (2007) method of gathering and weighting collocates based on distance in the dependency graph for the verbs (`pl2`). Herdağdelen and Baroni (2009) simply use a window-based approach for gathering collocates, which we reimplemented as a simpler way of capturing distributional similarity. The resulting features are named `w1` and `w2`.



His aunt bakes daily luscious cake
"his aunt bakes luscious cake every day"

`w12`	Seine$_{w_2,w_1}$ Tante$_{w_2,w_1}$ $_{w_2}$täglich$_{w_1}$
`lem12`	sein$_{w_2,w_1}$ Tante$_{w_2,w_1}$ $_{w_2}$täglich$_{w_1}$
`rel`	↑OBJA
`sat`	w_2ADV:täglich w_1AMOD:lecker w_1NSUBJ:Tante
`triples`	w_1 ↑OBJA w_2
`relsat`	↑OBJA/w_2ADV:täglich ↑OBJA/w_1AMOD:lecker ↑OBJA/w_1NSUBJ:Tante

Due to the short path between w_1 and w_2, the `triples` and `rel` features are not very different in the example. In case of more complicated constructions, the `triples` approach would yield multiple simpler features whereas `rel` would yield one single complex string.

Figure 1: Kinds of features

5 Count Transformations

It is a well-known fact in distributional semantics that raw observation counts for context items (be they elements surrounding single word occurrences or elements extracted from the occurences of two words together) are incomparable for different target words/target pairs (since their frequency can differ) as well as for different context items. As a result, researchers have proposed different approaches to produce transformed vectors using more sophisticated association statistics (see Dumais, 1991, Weeds et al., 2004, Turney and Pantel, 2010, *inter alia*).

In our case, we implemented L_1 normalization (which normalizes for target word frequency), a conservative estimate for pointwise mutual information (which normalizes for the frequencies of both target word and feature), and the G^2 log-likelihood measure of Dunning (1993), which gives significance scores (i.e., numbers that invariably grow both with target and feature frequency, even if the association strength – the relation between actual occurrences and those that would be expected when assuming no association – is constant). In both cases, very frequent features would be emphasized in comparison to medium- and low-frequency features.

In the realm of supervised learning, an additional choice has to be made among learning methods that can classify words or word pairs using large feature vectors – most commonly using nearest-neighbour classification (Nakov and Kozareva, 2011), using custom kernels in support vector classification (Ó Séaghdha and Copestake, 2009; Turney, 2008), or by using appropriate techniques to represent the feature vectors in linear classification.

In comparison to the former methods, linear classification scales better with the number of examples (where nearest-neigbour and kernel-based techniques both show strongly superlinear behaviour) and would be the method of choice for large-scale classification.

Herdağdelen and Baroni (2009) propose to map the values computed by association statistics by computing mean and standard deviation of each feature and mapping the range $[\mu - 2\sigma, \mu + 2\sigma]$ of association scores for that feature (seen over the values of that feature for all target pairs) to the range $[0, 1]$ in the input for the classifier, clamping values outside that range to 0 or 1, respectively.

Unfortunately, the approach proposed by Herdağdelen and Baroni has the property that an association score of 0 is mapped to a non-zero feature value for the classifier, which means that feature vectors are no longer sparse (i.e., instead of only storing non-zero values for context items that are informative, values for all context items have to be processed).

To keep the sparsity of the transformed counts, we always use 0 as the lower bound of the mapping (such that zero values stay zero values). In addition to the Herdagdelen and Baroni's mean/variance-based threshold, we investigated the following possibilities for fixing the upper bound:

- **MI scale**: use a constant upper bound of 1 on (a conservative estimate of) the pointwise mutual information.[2]

[2] To yield a conservative MI estimate, we use the discounting factor introduced by Pantel and Lin (2002). The pointwise mutual information value normalizes the frequency of both words of a pair, hence all mutual information values are on a common scale. A threshold of 1 in this case corresponds to two items oc-

baselines/single features	Acc	MacroF
random	0.463	0.090
telic-artifact	0.663	0.080
w12/L_1-norm/AMIS	0.677	0.181
w12/L_1-norm/SVM$_{\text{acc}}$	0.715	0.212
w12/L_1-norm/SVM$_{\text{F}}$	0.674	0.120
w12/L_1-norm	0.715	0.212
lem12/G^2-quant	0.703	0.204
rel/L_1-quant	0.722	0.154
sat/L_1-norm	0.700	0.185
triples/L_1-quant	**0.741**	0.192
triples/G^2-norm	0.739	**0.212**
relsat/L_1-quant	0.698	0.154
attr1+pl2/MI-thr	0.800	0.460
w1+w2/MI-thr	0.807	0.468
GermaNet, no combination	0.846	0.450
GermaNet, degree=2	**0.851**	**0.516**

Table 1: Trivial and single-feature baselines (using SVM-acc unless noted otherwise)

- **norm**: use a value based on mean and standard deviation of the occurring values for one given feature ($\mu + 2\sigma$).
- **quant**: use a fixed quantile (99%) of all values for a feature for the upper bound of the mapping interval.

In addition, to mapping feature values onto the unit interval $[0, 1]$, we investigated the usefulness of making the features binary-valued by mapping all values lower than the threshold to zero. While intuitively a continuous-valued feature should be more informative, the high dimensionality of the feature space may mean that noisy feature extraction ultimately leads to a worse model in the continuous-feature case.

6 Results and Discussion

Because of the skewed distribution, it is useful to look not only at the overall accuracy (Acc) but also at the macro-average of the F-measure of all relations (MacroF). The macro-averaged F-measure reflects the ability of the system to recognize all relations since it weighs all relation types equally, instead of (implicitly) weighting by token count where under-predicting rare relation types normally yields a higher accuracy. As is evident from table 1, the accuracy baseline for the most frequent label (*telic-artifact*) is already quite high.

combinations	Acc	MacroF
triples/G^2-norm	**0.739**	**0.212**
triples+w12/G^2-norm	0.733	0.206
triples+rel/G^2-norm	0.725	0.190
triples+sat/G^2-norm	0.738	0.200
triples+relsat/G^2-norm	0.729	0.184
triples+w1+w2/MI-thr	0.816	0.469
triples+attr1+pl2/MI-thr	0.807	0.431
GermaNet	0.851	**0.516**
GermaNet+triples/G^2-norm	0.853	0.482
GermaNet+triples/MI-thr	0.855	0.484
GermaNet+w12/G^2-norm	0.855	0.496
GermaNet+w12/MI-thr	**0.858**	0.510
GWN+w12+triples/G^2-norm	0.852	0.462
GWN+w12+triples/MI-thr	0.849	0.478
GermaNet+w1+w2/MI-thr	0.828	0.496

Table 2: Combination results (using SVMacc)

Looking at results with various scaling methods and learners on single features (table 1), we found that the SVM$_{\text{acc}}$ learner consistently yields better accuracy and macro-averaged F-measure than the other two learners. For the weighting functions, we found that none of the measures was consistently better than the others; results for the single features in table 1 are reported for a weighting function that works best for either accuracy or macro-averaged F-measure using. (For space reasons, table 1 shows numbers only for the w12 feature and L_1-norm scaling; other features and settings show a similar relation between the scores for different learners).

As in the investigation by Ó Séaghdha and Copestake (2007), dependency triples from the path between the two target words are the most effective feature representation and yields both the greatest accuracy value (with L_1 scaling and quantile-based setting of thresholds) and the greatest F-measure macroaverage (with G^2 scaling and setting of thresholds based on average and standard deviation). Combination of the `triples` feature with

curring together about $\exp(1) \approx 2.7$ times as often as would be expected from the marginal distribution for that co-occurrence relation.

	agentive	beh-anim	beh-artif	beh-body	beh-env	grooming	location	telic-artif	telic-body	telic-role
count	14	94	13	2	5	17	12	425	24	35
w12	0.105	0.513	0.000	0.000	0.000	0.000	0.154	0.834	0.214	0.255
triples	0.125	0.601	0.000	0.000	0.000	0.000	0.153	0.853	0.153	0.238
attr1+pl2	0.333	0.826	**0.258**	0.000	0.000	**0.500**	0.421	0.874	0.636	0.754
w1+w2	0.385	0.834	0.222	0.000	0.000	0.400	0.571	0.877	0.619	0.767
GermaNet	**0.480**	**0.859**	0.190	0.000	0.000	0.451	**0.636**	0.909	**0.773**	0.857
GWN+w12	0.400	0.857	0.133	0.000	**0.333**	0.384	0.600	**0.916**	0.600	**0.873**

Table 3: Results by relation

other features based on paired co-occurrences does not lead to further improvements, especially with those features that also express information from the dependency path (`rel`,`relsat`).

In comparison, the accuracy of the GermaNet hypernyms feature (which includes combinations of the hypernyms of first and second word) is much higher than the versions that do not make use of hand-crafted taxonomic knowledge, which is surprising since it uses only taxonomic and no relational information. The pairwise feature combination for GermaNet features yields another small improvement over these already very good results. Distributional information on single words, both the strictly window-based `w1+w2` feature and the one that is based on more elaborated distributional modeling (`attr1+pl2`) show quite good results that show further (but relatively small) improvements when combined with the `triples` feature.

The importance of taxonomic (or, alternatively, distributional semantic) information for the task proposed here - namely, the supervised classification of qualia-like relations - partly mirrors results for the supervised classification of relations between nominals, where Ó Séaghdha and Copestake (2007) find that their best system for distributional similarity based on the BNC performs at about the same level as a (somewhat simpler) approach using WordNet-based classification (Ó Séaghdha, 2007), with only much more sophisticated approaches such as the one of Ó Séaghdha and Copestake (2009), which also makes use of a considerably larger textual basis to improve results over the level of the WordNet-based approach.

Another reason for the importance of taxonomic information in this task may lie in the fact that the different relations have relatively strong selectional restrictions (for animate objects, roles/professions, body parts, or artifacts on the noun side, and certain types of actions or events on the verb side).

Looking at the results for each relation in table 3, we see that both *telic-artifact* and *behaviour-animate*, the two relations with the largest counts, are classified quite reliably, while *behaviour-bodypart* and *behaviour-environment*, the two relations with very few examples, are never found by the system. Among the other relations, taxonomical information for nouns and verbs seems to be instrumental for adequate classification of the *grooming* relation and possibly also for *location*, *telic-bodypart* and *telic-role*.

7 Summary

In this paper, we have presented a dataset containing cross-part-of-speech relations between concrete nouns and human verb associates and demonstrated a state-of-the-art approach for the supervised multiclass classification of the qualia relations in this dataset.[3] Our results show that taxonomic information from GermaNet is much superior to all other features, while corpus-based dependency triples are still visibly superior to shallow surface-based features.

Important questions for future research would include a more direct comparison to other languages (ideally using a similar data set and information sources) to tease apart the influences of word order, taxonomic organization, and data sparsity, respectively.

[3] The dataset and future corrected/improved versions, are available on request. Please feel free to send an email to the author if you want to use it or produce a create a version for another language.

Acknowledgements The work described in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (SFB) 833. The author gratefully acknowledges Melike Heubach's help with the annotation of the dataset.

References

Bergsma, S., Lin, D., and Goebel, R. (2008). Discriminative learning of selectional preference from unlabeled text. In *EMNLP 2008*.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL-1999*.

Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the Global WordNet Conference*.

Cassel, S. (2009). MaltParser and LIBLINEAR - transition-based dependency parsing with linear classification for feature model optimization. Master's thesis, Uppsala University.

Chaffin, R. and Herrmann, D. J. (1987). Relation element theory: A new account of the representation and processing of semantic relations. In Gorfein, D. S., editor, *Memory and Learning: the Ebbinghaus Centennial Conference*. Erlbaum.

Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proc. EMNLP 2004*.

Cimiano, P. and Wenderoth, J. (2005). Automatically learning qualia structures from the web. In *Proceedings of the ACL'05 Workshop on Deep Lexical Acquisition*.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Erk, K., Padó, S., and Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–761.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *NAACL-HLT 2003*.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., and Turney, P. (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.

Hare, M., Jones, M., Thomson, C., Kelly, S., and McRae, K. (2009). Activating event knowledge. *Cognition*, 111:151–167.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING 92)*.

Hearst, M. (1998). Automated discovery of wordnet relations. In *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge (MA), USA.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *SemEval 2010*.

Henrich, V. and Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.

Herdağdelen, A. and Baroni, M. (2009). BagPack: A general framework to represent semantic relations. In *ACL09 Workshop on Geometric Models of Natural Language Semantics (GEMS09)*.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Katrenko, S. and Adriaans, P. (2008a). Qualia structures and their impact on the concrete noun categorization task. In *ESSLLI 2008 workshop on Distributional Lexical Semantics*.

Katrenko, S. and Adriaans, P. (2008b). Semantic types of some generic relation arguments: Detection and evaluation. In *ACL/HLT 2008*.

Katrenko, S. and Adriaans, P. (2010). Finding constraints for semantic relations via clustering. In *CLIN 2010*.

Kurc, R. and Piasecki, M. (2008). Automatic acquisition of wordnet relations by the morpho-syntactic patterns extracted from the corpora in Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology - Third International Symposium Advances in Artificial Intelligence and Applications (IMCSIT 2008)*.

Lenci, A., Calzolari, N., and Zampolli, A. (2003). SIMPLE: Plurilingual semantic lexicons for natural language processing. *Linguistica Computatazionale*, 16–17:323–352.

McRae, K., Cree, G., Seidenberg, M., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behaviour Research Methods*, 37:547–559.

Melinger, A., Schulte im Walde, S., and Weber, A. (2006). Characterizing response types and revealing noun ambiguity in german association norms. In *Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41:197–229.

Miyao, Y. and Tsujii, J. (2002). Maximum entropy estimation for feature forests. In *HLT 2002*.

Morris, J. and Hirst, G. (2004). Non-classical lexical semantic relations. In *HLT/NAACL Workshop on Computational Lexical Semantics*.

Nakov, P. and Kozareva, Z. (2011). Combining relational and attributional similarity for semantic relation classification. In *RANLP 2011*.

Ó Séaghdha, D. and Copestake, A. (2009). Using lexical and relational similarity to classify semantic relations. In *EACL 2009*.

Ó Séaghdha, D. (2007). Annotating and learning compound noun semantics. In *Proceedings of the ACL07 Student Research Workshop*.

Ó Séaghdha, D. (2010). Latent variable models of selectional preference. In *ACL 2010*.

Ó Séaghdha, D. and Copestake, A. (2007). Co-occurrence contexts for noun compound interpretation. In *ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*.

Ó Séaghdha, D. and Copestake, A. (2009). Using lexical and relational similarity to classify semantic relations. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, pages 613–619.

Pantel, P. and Pennachiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *COLING/ACL 2006*.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Regneri, M. (2006). VerbOzean - maschinelles Lernen von semantischen Relationen zwischen deutschen Verben. Bachelorarbeit, Universität des Saarlandes.

Rüd, S. and Zarcone, A. (2011). Covert events and qualia structures for German verbs. In *Metonymy 2011*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *COLING 2008*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *NIPS 2005*.

Tjong Kim Sang, E. and Hofmann, K. (2009). Lexical patterns or dependency patterns: Which is better for hypernym extraction. In *CoNLL-2009*.

Turney, P. (2008). A uniform approach to analogies, synonyms, antonyms and associations. In *Coling 2008*.

Turney, P. and Littman, M. (2003). Learning analogies and semantic relations. Technical Report ERB-1103, National Research Council, Institute for Information Technology.

Turney, P. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Versley, Y. (2007). Using the Web to resolve coreferent bridging in German newspaper text. In *Proceedings of GLDV-Frühjahrstagung 2007*, Tübingen. Narr.

Versley, Y., Beck, A. K., Hinrichs, E., and Telljohann, H. (2010). A syntax-first approach to high-quality morphological analysis and lemma disambiguation for the TüBa-D/Z treebank. In *Proceedings of the 9th Conference on Treebanks and Linguistic Theories (TLT9)*.

Versley, Y. and Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. In *Proceedings of the 7th Web as Corpus Workshop (WAC-7)*, pages 44–52.

Verspoor, C. M. (1997). *Contextually-Dependent Lexical Semantics*. PhD thesis, University of Edinburgh.

Vigliocco, G., Vinson, D., Lewis, W., and Garrett, M. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48:422–488.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterizing measures of lexical distributional similarity. In *CoLing 2004*.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., and Sauri, R. (2006). Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proc. 7th SIGdial Workshop on Discourse and Dialogue*.

Yamada, I. and Baldwin, T. (2004). Automatic discovery of telic and agentive roles from corpus data. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18)*.