Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts

Alessandra Zarcone, Jason Utt
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{zarconaa,uttjn}@ims.uni-stuttgart.de

Sebastian Padó
Institut für Computerlinguistik
Universität Heidelberg
pado@cl.uni-heidelberg.de

Abstract

Logical metonymies (*The student finished the beer*) represent a challenge to compositionality since they involve semantic content not overtly realized in the sentence (covert events → *drinking the beer*). We present a contrastive study of two classes of computational models for logical metonymy in German, namely a probabilistic and a distributional, similarity-based model. These are built using the SDEWAC corpus and evaluated against a dataset from a self-paced reading and a probe recognition study for their sensitivity to thematic fit effects via their accuracy in predicting the correct covert event in a metonymical context. The similarity-based models allow for better coverage while maintaining the accuracy of the probabilistic models.

1 Introduction

Logical metonymies (*The student finished the beer*) require the interpretation of a *covert event* which is not overtly realized in the sentence (→ *drinking the beer*). Logical metonymy has received much attention as it raises issues that are relevant to both theoretical as well as cognitive accounts of language.

On the theoretical side, logical metonymies constitute a challenge for theories of compositionality (Partee et al., 1993; Baggio et al., in press) since their interpretation requires additional, inferred information. There are two main accounts of logical metonymy: According to the *lexical* account, a type clash between an event-subcategorizing verb (*finish*) and an entity-denoting object (*beer*) triggers the recovery of a covert event from complex lexical entries, such as qualia structures (Pustejovsky, 1995). The *pragmatic* account of logical metonymy suggests that covert events are retrieved through post-lexical inferences triggered by our world knowledge and communication principles (Fodor and Lepore, 1998; Cartson, 2002; De Almeida and Dwivedi, 2008).

On the experimental side, logical metonymy leads to higher processing costs (Pylkkänen and McElree, 2006; Baggio et al., 2010). As to covert event retrieval, it has been found that verbs cue fillers with a high thematic fit for their argument positions (e.g. $arrest \xrightarrow{agent} cop$, (Ferretti et al., 2001)) and that verbs and arguments combined cue fillers with a high thematic fit for the remaining argument slots (e.g. $\langle journalist, check \rangle \xrightarrow{patient} spelling$ but $\langle mechanic, check \rangle \xrightarrow{patient} car$ (Bicknell et al., 2010). The interpretation of logical metonymy is also highly sensitive to context (e.g. $\langle confectioner, begin, icing \rangle \xrightarrow{covertevent} spread$ but $\langle child, begin, icing \rangle \xrightarrow{covertevent} eat$ (Zarcone and Padó, 2011; Zarcone et al., 2012). It thus provides an excellent test bed for cognitively plausible computational models of language processing.

We evaluate two classes of computational models for logical metonymy. The classes represent the two main current approaches in lexical semantics: *probabilistic* and *distributional* models. Probabilistic models view the interpretation as the assignment of values to random variables. Their advantage is that they provide a straightforward way to include context, by simply including additional random variables. However, practical estimation of complex models typically involves independence assumptions, which

In: R. Levy & D. Reitter (Eds.), *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 70–79, Montréal, Canada, June 7, 2012.

may or may not be appropriate, and such models only take first-order co-occurrence into account[1]. In contrast, distributional models represent linguistic entities as co-occurrence vectors and phrase interpretation as a vector similarity maximization problem. Distributional models typically do not require any independence assumptions, and include second-order co-occurrences. At the same time, how to integrate context into the vector computation is essentially an open research question (Mitchell and Lapata, 2010).

In this paper, we provide the first (to our knowledge) distributional model of logical metonymy by extending the context update of Lenci's ECU model (Lenci, 2011). We compare this model to a previous probabilistic approach (Lapata and Lascarides, 2003a; Lapata et al., 2003b). In contrast to most experimental studies on logical metonymy, which deal with English data (with the exception of Lapata et al. (2003b)), we focus on German. We estimate our models on a large web corpus and evaluate them on a psycholinguistic dataset (Zarcone and Padó, 2011; Zarcone et al., 2012). The task we use to evaluate our models is to distinguish covert events with a high typicality / thematic fit (e.g. *The student finished the beer* $\longrightarrow$ *drinking*) from low typicality / thematic fit covert events ($\longrightarrow$ *brewing*).

2 Probabilistic models of logical metonymy

Lapata et al. (2003b; 2003a) model the interpretation of a logical metonymy (e.g. *The student finished the beer*) as the joint distribution $P(s, v, o, e)$ of the variables s (the subject, e.g. *student*), v (the metonymic verb, e.g. *finish*), o (the object, e.g. *beer*), e (the covert event, *drinking*).

This model requires independence assumptions for estimation. We present two models with different independence assumptions.

[1]This statement refers to the simple probabilistic models we consider, which are estimated directly from corpus co-occurrence frequencies. The situation is different for more complex probabilistic models, for example generative models that introduce latent variables, which can amount to clustering based on higher-order co-occurrences, as in, e.g., Prescher et al. (2000).

2.1 The SOV_p model

Lapata et al. develop a model which we will refer to as the SOV_p *model*.[2] It assumes a generative process which first generates the covert event e and then generates all other variables based on the choice of e:

$$P(s, v, o, e) \approx P(e)\ P(o|e)\ P(v|e)\ P(s|e)$$

They predict that the selected covert event $\hat{e}$ for a given context is the event which maximizes $P(s, v, o, e)$:

$$\hat{e} = \arg\max_e P(e)\ P(o|e)\ P(v|e)\ P(s|e)$$

These distributions are estimated as follows:

$$\hat{P}(e) = \frac{f(e)}{N}, \quad \hat{P}(o|e) = \frac{f(e \overset{o}{\leftarrow} o)}{f(e \overset{o}{\leftarrow} \cdot)},$$
$$\hat{P}(v|e) = \frac{f(v \overset{c}{\leftarrow} e)}{f(\cdot \overset{c}{\leftarrow} e)}, \quad \hat{P}(s|e) = \frac{f(e \overset{s}{\leftarrow} s)}{f(e \overset{s}{\leftarrow} \cdot)},$$

where N is the number of occurrences of full verbs in the corpus; $f(e)$ is the frequency of the verb e; $f(e \overset{o}{\leftarrow} \cdot)$ and $f(e \overset{s}{\leftarrow} \cdot)$ are the frequencies of e with a direct object and subject, respectively; and $f(e \overset{c}{\leftarrow} \cdot)$ is number of times e is the complement of another full verb.

2.2 The SO_p model

In Lapata et al.'s covert event model, v, the metonymic verb, was used to prime different choices of e for the same object (*begin book*$\longrightarrow$ writing; *enjoy book*$\longrightarrow$ reading). In our dataset (Sec. 4), we keep v constant and consider e only as a function of s and o. Thus, the second model we consider is the SO_p *model* which does not consider v:

$$P(s, v, o, e) \approx P(s, o, e) \approx P(e)\ P(o|e)\ P(s|e)$$

Again, the preferred interpretation $\hat{e}$ is the one that maximizes $P(s, v, o, e)$:

$$\hat{e} = \arg\max_e P(e)\ P(o|e)\ P(s|e)$$

[2]In Lapata et al. (2003b; 2003a), this model is called the *simplified* model to distinguish it from a *full* model. Since the full model performs worse, we do not include it into consideration and use a more neutral name for the simplified model.

3 Similarity-based models

3.1 Distributional semantics

Distributional or vector space semantics (Turney and Pantel, 2010) is a framework for representing word meaning. It builds on the Distributional Hypothesis (Harris, 1954; Miller and Charles, 1991) which states that words occurring in similar contexts are semantically similar. In distributional models, the meaning of a word is represented as a vector whose dimensions represent features of its linguistic context. These features can be chosen in different ways; popular choices are simple words (Schütze, 1992) or lexicalized dependency relations (Lin, 1998; Padó and Lapata, 2007). Semantic similarity can then be approximated by vector similarity using a wide range of similarity metrics (Lee, 1999).

3.1.1 Distributional Memory

A recent multi-purpose framework in distributional semantics is Distributional Memory (DM, Baroni and Lenci (2010)). DM does not immediately construct vectors for words. Instead, it extracts a three-dimensional tensor of weighted *word-link-word* tuples each of which is mapped onto a score by a function $\sigma\colon \langle w_1\ l\ w_2\rangle \rightarrow \mathbb{R}^+$. For example, $\langle$*pencil obj use*$\rangle$ has a higher weight than $\langle$*elephant obj use*$\rangle$. The set of links can be defined in different ways, yielding various DM instances. Baroni and Lenci present DepDM (mainly syntactic links such as $subj\_tr$), LexDM (strongly lexicalized links, e.g., $such\_as$), or TypeDM (syntactic and lexicalized links).[3]

The benefit of the tensor-based representation is that it is general, being applicable to many tasks. Once a task is selected, a dedicated semantic space for this task can be generated efficiently from the tensor. For example, the *word by link-word* space ($W_1 \times LW_2$) contains vectors for the words w_1 whose dimensions are labeled with $\langle l, w_2\rangle$ pairs. The *word-word by link* space ($W_1W_2 \times L$) contains co-occurrence vectors for word pairs $\langle w_1, w_2\rangle$ whose dimensions are labeled with l.

3.2 Compositional Distributional Semantics

Probabilistic models can account for compositionality by estimating conditional probabilities. Compositionality is less straightforward in a similarity-based distributional model, because similarity-based distributional models traditionally model meaning at word level. Nevertheless, the last years have seen a wave of distributional models which make progress at building compositional representations of higher-level structures such as noun-adjective or verb-argument combinations (Mitchell and Lapata, 2010; Guevara, 2011; Reddy et al., 2011).

3.2.1 Expectation Composition and Update

Lenci (2011) presents a model to predict the degree of thematic fit for verb-argument combinations: the Expectation Composition and Update (ECU) model. More specifically, the goal of ECU is explain how the choice of a specific subject for a given verb impacts the semantic expectation for possible objects. For example, the verb *draw* alone might have fair, but not very high, expectations for the two possible objects *landscape* and *card*. When it is combined with the subject *painter*, the resulting phrase *painter draw* the expectation for the object *landscape* should increase, while it should drop for *card*.

The idea behind ECU is to first compute the verb's own expectations for the object from a TypeDM $W_1 \times LW_2$ matrix and then update it with the subject's expectations for the object, as mediated by the TypeDM *verb* link type.[4] More formally, the verb's expectations for the object are defined as

$$EX_V(v) = \lambda o.\, \sigma(\langle v\ \textit{obj}^{-1}\ o\rangle)$$

The subject's expectations for the object are

$$EX_S(s) = \lambda o.\, \sigma(\langle s\ \textit{verb}\ o\rangle)$$

And the updated expectation is

$$EX_{SV}(s, v) = \lambda o.\, EX_V(v)(o) \circ EX_S(s)(o)$$

where $\circ$ is a composition operation which Lenci instantiates as sum and product, following common practice in compositional distributional semantics (Mitchell and Lapata, 2010). The product composition approximates a conjunction, promoting objects that are strongly preferred by both verb and subject. It is, however, also prone to sparsity problems as well

[3] l^{-1} is used to denote the inverse link of l (i.e., exchanging the positions of w_1 and w_2).

[4] In DM, *verb* directly connects the subject and the object of transitive verb instances, e.g $\langle$*marine verb gun*$\rangle$.

shortcomings of the scoring function σ. The sum composition is more akin to a disjunction where it suffices that an object is strongly preferred by either the verb or the subject.

It would be possible to use these scores as direct estimates of expectations, however, since EX_{SV} contains three lexical variables, sparsity is a major issue. ECU thus introduces a distributional generalization step. It only uses the updated expectations to identify the 20 most expected nouns for the object position. It then determines the prototype of the updated expectations as the centroid of their $W_1 \times LW_2$ vectors. Now, the thematic fit for any noun can be computed as the similarity of its vector to the prototype.

Lenci evaluates ECU against a dataset from Bicknell et al. (2010), where objects (e.g. *spelling*) are matched with a high-typicality subject-verb combinations (e.g. $\langle$*journalist, check*$\rangle$ - high thematic fit) and with a low-typicality subject-verb combination (e.g. $\langle$*mechanic, check*$\rangle$ - low thematic fit). ECU is in fact able to correctly distinguish between the two contexts differing in thematic fit with the object.

3.3 Cognitive relevance

Similarity-based models build upon the Distributional Hypothesis, which, in its strong version, is a cognitive hypothesis about the form of semantic representations (Lenci, 2008): the distributional behavior of a word reflects its semantic behavior but is also a direct correlate of its semantic content at the cognitive level. Also, similarity-based models are highly compatible with known features of human cognition, such as graded category membership (Rosch, 1975) or multiple sense activation (Erk, 2010). Their cognitive relevance for language has been supported by studies of child lexical development (Li et al., 2004), category-related deficits (Vigliocco et al., 2004), selectional preferences (Erk, 2007), event types (Zarcone and Lenci, 2008) and more (see Landauer et al. (2007) and Baroni and Lenci (2010) for a review).

3.4 Modeling Logical Metonymy with ECU

3.4.1 Logical Metonymy as Thematic Fit

The hypothesis that we follow in this paper is that the ECU model can also be used, with modifications, to predict the interpretation of logical metonymy. The underlying assumption is that the interpretation of logical metonymy is essentially the recovery of a covert event with a maximal thematic fit (high-typicality) and can thus make use of ECU's mechanisms to treat verb-argument composition. Strong evidence for this assumption has been found in psycholinguistic studies, which have established that thematic fit dynamically affects processing, with online updates of expectations for typical fillers during the incremental processing of linguistic input (see McRae and Matsuki (2009) for a review). Thus, we can hope to transfer the benefits of similarity-based models (notably, high coverage) to the interpretation of logical metonymy.

3.4.2 Extending ECU

The ECU model nevertheless requires some modifications to be applicable to logical metonymy. Both the entity of interest and the knowledge sources change. The entity of interest used to be the object of the sentence; now it is the covert event, which we will denote with e. As for knowledge sources, there are three sources in logical metonymy. These are (a), the subject (compare *the author began the beer* and *the reader began the book*)); (b), the object *the reader began the book* vs. *the reader began the sandwich*); and (c), the metonymic verb (compare *Peter began the report* vs. *Peter enjoyed the report*).

The basic equations of ECU can be applied to this new scenario as follows. We first formulate three basic equations that express the expectations of the covert event given the subject, object, and metonymic verb individually. They are all derived from direct dependency relations in the DM tensor (e.g., the novel metonymic verb–covert event relation from the verbal complement relation):

$$EX_S(s) = \lambda e.\, \sigma(\langle s \textit{ subj } e\rangle)$$
$$EX_O(o) = \lambda e.\, \sigma(\langle o \textit{ obj } e\rangle)$$
$$EX_V(v) = \lambda e.\, \sigma(\langle v \textit{ comp}^{-1} \; e\rangle)$$

To combine (or update) these basic expectations into a final expectation, we propose two variants:

ECU SOV In this model, we compose all three expectations:

$$EX_{SOV}(s,v,o) = \lambda e.\, EX_S(s)(e) \circ EX_O(o)(e) \circ EX_V(v)(e)$$

	CE	
	high thematic fit	low thematic fit
Der *Konditor* begann, die *Glasur* The baker started the icing	**aufzutragen**. **to spread**.	**zu essen**. **to eat**.
Das *Kind* begann, die *Glasur* The child started the icing	**zu essen**. **to eat**.	**aufzutragen**. **to spread**.

Table 1: Example materials for the self-paced reading and probe recognition studies

We will refer to this model as SOV_{Σ} when the composition function is sum, and as the SOV_{Π} model when the composition function is product.

ECU SO Analogous to the SO probabilistic model, this model abstracts away from the metonymic verb. We assume most information about an event to be determined by the subject and object:

$$EX_{SO}(n, n') = \lambda e.\, EX_S(n)(e) \circ EX_O(n')(e)$$

After the update, the prototype computation proceeds as defined in the original ECU.

We will refer to this model as SO_{Σ} when the composition function is sum, and as the SO_{Π} model when the composition function is product.

4 Experimental Setup

We evaluate the probabilistic models (Sec. 2) and the similarity-based models (Sec. 3) on a dataset constructed from two German psycholinguistic studies on logical metonymy. One study used self-paced reading and the second one probe recognition.

Dataset The dataset we use is composed of 96 sentences. There are 24 sets of four $\langle s, v, o, e\rangle$ tuples, where s is the object, v the metonymic verb, o the object and e the covert event. The materials are illustrated in Table 1. As can be seen, all tuples within a set share the same metonymic verb and the same object. Each of the two subject e is matched once with a high-typicality covert event and once with a low-typicality covert event. This results in 2 high-typicality tuples and 2 low-typicality tuples in each set. Typical events (e) were elicited by 20 participants given the corresponding object o, subjects were elicited by 10 participants as the prototypical agents subjects for each e, o combination.

The experiments yielded a main effect of typicality on self-paced reading times (Zarcone and Padó, 2011) and on probe recognition latencies (Zarcone et al., 2012): typical events involved in logical metonymy interpretation are read faster and take longer to be rejected as probe words after sentences which evoke them. The effect is seen early on (after the patient position in the self-paced reading and at short ISI for the probe recognition), suggesting that knowledge of typical events is quickly integrated in processing and that participants access a broader pool of knowledge than what has traditionally been argued to be in the lexical entries of nouns (Pustejovsky, 1995). The finding is in agreement with results of psycholinguistic studies which challenge the very distinction between world knowledge and linguistic knowledge (Hagoort et al., 2004; McRae and Matsuki, 2009).

DM for German Since DM exists only for English, we constructed a German analog using the 884M word SDEWAC web corpus (Faaß et al., 2010) parsed with the MATE German dependency parser (Bohnet, 2010).

From this corpus, we extract 55M instances of simple syntactic relations (*subj_tr, subj_intr, obj, iobj, comp, nmod*) and 104M instances of lexicalized patterns such as *noun–prep–noun* e.g. ⟨*Recht auf Auskunft*⟩ (⟨*right to information*⟩), or *adj–noun-(of)-noun* such as ⟨*strittig Entscheidung Schiedsrichter*⟩ (⟨*contested decision referee*⟩). These lexicalized patterns make our model roughly similar to the English TypeDM model (Sec. 3.1.1).

As for σ, we used local mutual information (LMI) as proposed by Baroni and Lenci (2010). The LMI of a triple is defined as $O_{w_1lw_2} \log(O_{w_1lw_2}/E_{w_1lw_2})$, where $O_{w_1lw_2}$ is the observed co-occurrence frequency of the triple and $E_{w_1lw_2}$ its expected co-occurrence frequency (under the assumption of independence). Like standard MI, LMI measures the informativity or surprisal of a co-occurrence, but

weighs it by the observed frequency to avoid the overestimation for low-probability events.

4.1 Task

We evaluate the models using a binary selection task, similar to Lenci (2011). Given a triple $\langle s, v, o\rangle$ and a pair of covert events e, e' (cf. rows in Tab. 1), the task is to pick the high-typicality covert event for the given triple: $\langle$*Chauffeur*, *vermeiden*, *Auto*$\rangle \rightarrow$ *fahren*/*reparieren* ($\langle$*driver*, *avoid*, *car*$\rangle \rightarrow$ *drive*/*repair*). Since our dataset consists of 96 sentences, we have 48 such contexts.

With the probabilistic models, we compare the probabilities $P(s, v, o, e)$ and $P(s, v, o, e')$ (ignoring v in the SO model). Analogously, for the similarity-based models, we compute the similarities of the vectors for e and e' to the prototype vectors for the expectations $EX_{SOV}(s, v, o)$ and predict the one with higher similarity. For the simplified ECU SO model, we use $EX_{SO}(s, o)$ as the point of comparison.

4.2 Baseline

Following the baseline choice in Lapata et al. (2003b), we evaluated the probabilistic models against a baseline (B_p) which, given a $\langle s, v, o\rangle$ triplet (e.g. $\langle$*Chauffeur, vermeiden, Auto*$\rangle$), scores a "hit" if the $\hat{P}(e|o)$ for the high-typicality e is higher than the $\hat{P}(e|o)$ for the low-typicality e. The similarity-based models were evaluated against a baseline (B_s) which, given an $\langle s, v, o\rangle$ triplet (e.g. $\langle$*Chauffeur, vermeiden, Auto*$\rangle$), makes a correct prediction if the prototypical event vector for o has a higher thematic fit (i.e. similarity) with the high-typicality e than with the low-typicality e.

Since our dataset is counterbalanced – that is, each covert event appears once as the high-typicality event for a given object (with a congruent subject) and once as the low-typicality event – the baseline predicts the correct covert event in exactly 50% of the cases. Note, however, that this is not a *random* baseline: the choice of the covert event is made deterministically on the basis of the input parameters.

4.3 Evaluation measures

We evaluate the output of the model with the standard measures coverage and accuracy. *Coverage* is defined as the percentage of datapoints for which a model can make a prediction. Lack of coverage arises primarily from sparsity, that is, zero counts for co-occurrences that are necessary in the estimation of a model. *Accuracy* is computed on the covered contexts only, as the ratio of correct predictions to the number of predictions of the model. This allows us to judge the quality of the model's predictions independent of its coverage.

We also consider a measure that combines coverage and accuracy, *Backoff Accuracy*, defined as: $\mathit{coverage} \times \mathit{accuracy} + ((1 - \mathit{coverage}) \times 0.5)$. Backoff Accuracy emulates a backoff procedure: the model's predictions are adopted where they are available; for the remaining datapoints, it assumes baseline performance (in the current setup, 50%). The Backoff Accuracy of low-coverage models tends to degrade towards baseline performance.

We determine the significance of differences between models with a χ^2 test, applied to a 2×2 contingency matrix containing the number of correct and incorrect answers. Datapoints outside a model's coverage count half for each category, which corresponds exactly to the definition of Backoff Accuracy.

5 Results

The results are shown in Table 2. Looking at the probabilistic models, we find SO_p yields better coverage and better accuracy than SOV_p (Lapata's *simplified model*). It is worth noting the large difference in coverage, namely .75 as opposed to .44: The SOV_p model is unable to make a prediction for more than half of all contexts. This is due to the fact that many $\langle o, v\rangle$ combinations are unattested in the corpus. Even on those contexts for which the probabilistic SOV_p model can make a prediction, it is less reliable than the more general SO_p model (0.62 versus 0.75 accuracy). This indicates that, at least on our dataset, the metonymic verb does not systematically help to predict the covert event; it rather harms performance by introducing noisy estimates. As the lower half of the Table shows, the SOV_p model does not significantly outperform any other model (including both baselines B_p and B_s).

The distributional models do not have such coverage issues. The main problematic combination for the similarity model is $\langle$*Pizzabote hassen Pizza*$\rangle$ (i.e. $\langle$*Pizza delivery man hate pizza*$\rangle$) which is paired with the covert events *liefern (deliver)* and *backen (bake)*. The computation of ECU predictions for

	Probabilistic Models			Similarity-based Models				
	B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Accuracy	0.50	0.62	0.75	0.50	0.68	0.56	0.68	0.70
Coverage	1.00	0.44	0.75	1.00	0.98	0.94	0.98	0.98
Backoff Accuracy	0.50	0.55	**0.69**	0.50	0.68	0.56	0.68	**0.70**

		Probabilistic Models			Similarity-based Models				
		B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Prob.	B_p								
	SOV_p	-							
	SO_p	*	-						
Similarity	B_s	-	-	*					
	SOV_Σ	*	-	-	*				
	SOV_Π	-	-	-	-	-			
	SO_Σ	*	-	-	*	-	-		
	SO_Π	**	$*^\dagger$	-	**	-	$*^\dagger$	-	

Table 2: Results (above) and significance levels for difference in backoff accuracy determined by χ^2-test (below) for all probabilistic and similarity-based models (**: p<0.01, *: p≤0.05, -: p>0.05). For $*^\dagger$ ($SO_\Pi - SOV_p$ and $SO_\Pi - SOV_\Pi$) p was just above 0.05 (p=0.053).

this combination requires corpus transitive corpus constructions for *Pizzabote*, in the corpus it is only attested once as the subject of the intransitive verb *kommen* (*come*).

Among distributional models, the difference between SO and SOV is not as clear-cut as on the probabilistic side. We observe an interaction with the composition operation. Sum is less sensitive to complexity of updating: for sum models, the inclusion of the metonymic verb (SOV_Σ vs. SOV_Π) does not make any difference. On the side of the product models, there is a major difference similar to the one for the probabilistic models: SOV_Π is the worst model at near-baseline performance, and SO_Π is the best one. This supports our interpretation from above that the metonymic model introduces noisy expectations which, in the product model, have the potential of disrupting the update process.

Comparing the best models from the probabilistic and similarity-based classes (SO_p and SO_Π), we find that both significantly outperform the baselines. This shows that the subject contributes to the models with a significant improvement over the baseline models, which are only informed by the object. Their backoff accuracies do not significantly differ from one another, which is not surprising given the small size of our dataset, however, the similarity-based model outperforms the probabilistic model by 1% Backoff Accuracy. The two models have substantially different profiles: the accuracy of the probabilistic model is 5% higher (0.70 vs. 0.75); at the same time, its coverage is much lower. It covers only 75% of the contexts, while the distributional model SO_Π covers all but one (98%).

6 Discussion

As mentioned above, the main issue with the probabilistic models is coverage. This is due to the reliance of these models on first-order co-occurrence.

For example, probabilistic models cannot assign a probability to any of the triples ⟨*Dieb/Juwelier schmuggeln/schleifen Diamant*⟩ (⟨*thief/jeweler smuggle/cut diamond*⟩), since the subjects do not occur with either of the verbs in corpus, even though *Diamant* does occur as the object of both.

In contrast, the similarity-based models are able to compute expectations for these triples from second-order co-occurrences by taking into account other verbs that co-occur with *Diamant*. The ECU model is not punished by the extra context, as both *Dieb* and *Diamant* are associated with the verbs: *stehlen* (steal),

$EX_{SO}(\langle Chauffeur, Auto\rangle)$		$EX_{SO}(\langle Mechaniker, Auto\rangle)$	
fahren	*(drive)*	*bauen*	*(build)*
parken	*(park)*	*lassen*	*(let/leave)*
lassen	*(let/leave)*	*besitzen*	*(own)*
geben	*(give)*	*reparieren*	*(repair)*
sehen	*(see)*	*brauchen*	*(need)*
bringen	*(bring)*	*sehen*	*(see)*
steuern	*(steer)*	*benutzen*	*(use)*
halten	*(keep/hold)*	*stellen*	*(put)*

Table 3: Updated expectations in SO_{Π} for *Chauffeur* (*chauffeur*), *Mechaniker* (*mechanic*) and *Auto* (*car*).

rauben (thieve), *holen* (*get*), *entwenden* (*purloin*), *erbeuten* (*snatch*), *verkaufen* (*sell*), *nehmen* (*take*), *klauen* (*swipe*). We also note that these are typical events for a thief, which fits the intuition that *Dieb* is more predictive of the event than *Diamant*.

For both $\langle$*Chauffeur Auto*$\rangle$ and $\langle$*Mechaniker Auto*$\rangle$ the probabilistic model predicts *fahren* due to the high overall frequency of *fahren*.[5] The distributional model, however, takes the mutual information into account and is thus able to determine events that are more strongly associated with *Mechaniker* (e.g. *bauen, reparieren*, etc.) while at the same time discounting the uninformative verb *fahren*.

There are, however, items that all models have difficulty with. Three such cases are due to a frequency disparity between the high and low-typicality event. E.g. for $\langle$*Lehrerin Klausur benoten/schreiben*$\rangle$ ($\langle$*teacher exam grade/take*$\rangle$), *schreiben* occurs much more frequently than *benoten*. In the case of $\langle$*Schüler Geschichte lernen/schreiben*$\rangle$ ($\langle$*student story learn/write*$\rangle$), none of the models or baselines correctly assigned *lernen*. The probabilistic models are influenced by the very frequent *Geschichte schreiben* which is part of an idiomatic expression (*to write history*). On the other hand, the distributional models judge the *story* and *history* sense of the word to have the most informative events, e.g. *erzählen* (*tell*), *lesen* (*read*), *hören* (*hear*), *erfinden* (*invent*), and *studieren* (*study*), *lehren* (*teach*).

The baselines were able to correctly choose *auspacken* (*unwrap*) over *einpacken* (*wrap*) for $\langle$*Geburtstagskind Geschenk*$\rangle$ ($\langle$*birthday-boy/girl present*$\rangle$) while the models were not. The probabilistic models lacked coverage and were not able to make a prediction. For the distributional models, while both *auspacken* and *verpacken* (*wrap*) are highly associated with *Geschenk*, the most strongly associated actions of *Geburtstagskind* are extraordinarily diverse, e.g.: *bekommen* (*receive*), *sagen* (*say*), *auffuttern* (*eat up*), *herumkommandieren* (*boss around*), *ausblasen* (*blow out*). Neither of the events of interest though were highly associated.

[5]The combination *Mechaniker fahren* was seen once more often than *Mechaniker reparieren*.

7 Future Work

We see a possible improvement in the choice of the number of fillers, with which we construct the prototype vectors. A smaller number might lead to less noisy prototypes.

It has been shown (Bergsma et al., 2010) that the meaning of the prefix verb can be accurately predicted using the stem's vector, when compositionality applies. We suspect covert events that are prefix verbs to suffer from sparser representations than the vectors of their stem. E.g., *absaugen* (*vacuum off*) is much less frequent than the semantically nearly identical *saugen* (*vacuum*). Thus, by leveraging the richer representation of the stem, our distributional models could more likely assign the correct event.

8 Conclusions

We have presented a contrastive study of two classes of computational models, probabilistic and distributional similarity-based ones, for the prediction of covert events for German logical metonymies.

We found that while both model classes models outperform baselines which only take into account information coming from the object, similarity-based models rival and even outperform probabilistic models. The reason is that probabilistic models have to rely on first-order co-occurrence information which suffers from sparsity issues even in large web corpora. This is particularly true for languages like German that have a complex morphology, which tends to aggravate sparsity (e.g., through compound nouns).

In contrast, similarity-based models can take advantage of higher-order co-occurrences. Provided that some care is taken to identify reasonable vector composition strategies, they can maintain the accuracy of probabilistic models while guaranteeing higher coverage.

Acknowledgments

We would like to thank Alessandro Lenci, Siva Reddy and Sabine Schulte im Walde for useful feedback and discussion. The research for this paper has been funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the SFB 732 "Incremental specification in context" / project D6 "Lexical-semantic factors in event interpretation" at the University of Stuttgart.

References

Giosuè Baggio, Travis Chroma, Michiel van Lambalgen, and Peter Hagoort. 2010. Coercion and compositionality. *Journal of Cognitive Neuroscience*, 22(9):2131–2140.

Giosuè Baggio, Michiel van Lambalgen, and Peter Hagoort. in press. The processing consequences of compositionality. In *The Oxford Handbook of Compositionality*. Oxford University Press.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1–49.

Shane Bergsma, Aditya Bhargava, Hua He, and Grzegorz Kondrak. 2010. Predicting the semantic compositionality of prefix verbs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 293–303, Cambridge, MA, October. Association for Computational Linguistics.

Klinton Bicknell, Jeffrey L. Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Robyn Cartson. 2002. *Thoughts and utterances*. Blackwell.

Roberto G. De Almeida and Veena D. Dwivedi. 2008. Coercion without lexical decomposition: Type-shifting effects revisited. *Canadian Journal of Linguistics*, 53(2/3):301–326.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of ACL*, Prague, Czech Republic.

Katrin Erk. 2010. What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the workshop on Geometrical Models of Natural Language Semantics (GEMS)*, Uppsala, Sweden.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR as an Example of Morphological Evaluation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

T. R. Ferretti, K. McRae, and A. Hatherell. 2001. Integrating verbs, situation schemas and thematic role concept. *Journal of Memory and Language*, 44:516–547.

Jerry A. Fodor and Ernie Lepore. 1998. The emptiness of the lexicon: Reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry*, 29(2):269–288.

Emiliano Raul Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of IWCS-2011*, Oxford, UK.

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *Science*, 304:438–441.

Zelig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Mirella Lapata and Alex Lascarides. 2003a. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):263–317.

Mirella Lapata, Frank Keller, and Christoph Scheepers. 2003b. Intra-sentential context effects on the interpretation of logical metonymy. *Cognitive Science*, 27(4):649–668.

Lillian Lee. 1999. Measures of Distributional Similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, MA.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science. Special issue of the Italian Journal of Linguistics*, 20(1):1–31.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, Oregon.

Ping Li, Igor Farkas, and Brian MacWhinney. 2004. Early lexical development in a self-organizing neural network. *Neural Networks*, 17:1345–1362.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL*, pages 768–774, Montreal, QC.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3/6:1417–1429.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33:161–199, June.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. 1993. *Mathematical Methods in Linguistics*. Kluwer.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of COLING 2000*, Saarbrücken, Germany.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Liina Pylkkänen and Brian McElree. 2006. The syntax-semantic interface: On-line composition of sentence meaning. In *Handbook of Psycholinguistics*, pages 537–577. Elsevier.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP 2011*, Chiang Mai, Thailand.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787 –796.

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

Gabriella Vigliocco, David P. Vinson, William Lewis, and Merrill F. Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4):422–488.

Alessandra Zarcone and Alessandro Lenci. 2008. Computational models for event type classification in context. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. ELRA.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized event knowledge in logical metonymy resolution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 944–949, Austin, TX.

Alessandra Zarcone, Sebastian Padó, and Alessandro Lenci. 2012. Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Austin, TX.