Instance-based disambiguation of English *-ment* derivatives

MARIOS ANDREOU[1], LEA KAWALETZ[1], MAX KISSELEW[2], GABRIELLA LAPESA[2], SEBASTIAN PADO[2] & INGO PLAG[1]
([1]Heinrich-Heine-Universität Düsseldorf, [2]Universität Stuttgart)

One of the central problems in the semantics of derived words is polysemy. As shown by Lieber (in press), this kind of polysemy can be disambiguated in context. Recent work within the framework of Frame Semantics (Kawaletz and Plag, 2015; Plag, Andreou, and Kawaletz, to appear), however, has uncovered that context does not always fully determine the semantics of a given derived word.

In this paper, we tackle the problem of disambiguating newly derived words in context, based on the Distributional Semantics methodology (Firth, 1957). We use corpus-extracted representations to interpret deverbal *-ment* nominalizations that are either **event-denoting** (e.g. *assessment*) or **entity-denoting** (e.g. *pavement*).

Our dataset comprises low frequency *-ment* derivatives (59 types, 407 tokens). We chose low frequency derivatives since high frequency formations are often lexicalized and thus do not lend themselves readily to the kind of polysemous readings we are interested in. The 59 types are based on four verb classes: psych verbs (*annoy*), verbs of putting (*embed*), force verbs (*coerce*) and change of state verbs (*congeal*). Verb occurrences were extracted from corpora such as the Corpus of Contemporary American English, the Corpus of GlobalWeb-Based English, and WebCorp. In the following examples from our dataset, *emplacement* in (1) denotes an event, *bedragglement* in (2) denotes an entity, and *embrittlement* in (3) is ambiguous between an event and an entity:

(1) Event-denoting nominalization: In many places, **emplacement** of granite plutons is synchronous to volcanic eruptions. (Google Website 1995)
(2) Entity-denoting nominalization: I set down the scrap of doll's dress, a **bedragglement** of loose lace hem (COCA FIC 1999)
(3) Ambiguous reading: After 8 weeks of hydrolytic degradation, the nonwoven fabric was broken. There is an obvious **embrittlement** and cracking on the nonwoven fabric (Figure 6.5b). (Google ACAD 2014)

In distributional semantic modeling, co-occurrence information for all instances of a word is commonly conflated into a unique distributional vector (DSMs usually model word types, not tokens). This approach provides reliable semantic representations only if the target words are frequent enough. Our dataset, however, contains low-frequency words as well as attestation-specific annotations. Therefore, we adopt a token-based perspective on distributional modeling: We consider the sentence in which the disambiguation candidate occurs as a bag of contextual cues and employ machine learning to learn which cues are discriminative for the event vs. entity interpretation (e.g. temporal modifiers are associated with events, and physical descriptors with entities).

Figure 1 displays our experimental pipeline in terms of three macro-steps: distributional modeling, machine learning, and evaluation of the dataset described above.

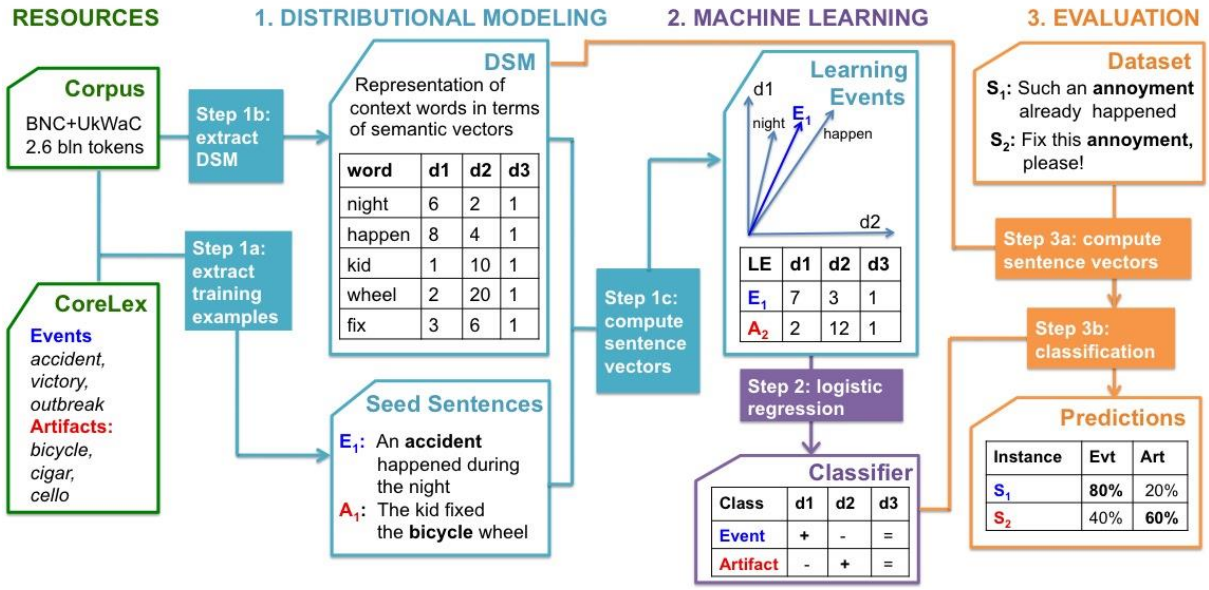

Figure 1: Instance-based disambiguation of *-ment* derivatives: pipeline

Macro-step 1 is the distributional modeling of the training data to be fed to the machine learning module (macro-step 2). To learn the discriminative features of entity vs. event nouns we need training examples and the corresponding bag-of-context-cues representation. We extract training examples from our source corpus (a concatenation of the British National Corpus and UkWaC, 2.6 billion tokens in total) as follows:

Step 1a: Extraction of the training examples from the source corpus. The source of the training set of prototypical event and entity nouns (our training *seeds*) is CoreLex (Buitelaar, 1998). Two sets of words which unambiguously denote events or entities were extracted from the CoreLex classes EVENT and ARTIFACT; we selected 196 items per class matched by frequency in the source corpus (min:101; mean:3.5k; max:127k). From the source corpus, we then extracted the sentences containing the target seeds and randomly sampled 100 sentences per seed word. These sentences represent the input of step 1c.

Step 1b: Extraction of a DSM from the source corpus. We extracted a DSM for inflected words with a very large vocabulary, including all words occurring more than 10 times in the source corpus. Applying a state-of-the-art DSM technique (Mikolov et al., 2013) we constructed a DSM with only 300 dimensions. Dimensionality reduction techniques produce more manageably sized vectors and improve the semantic representations by inferring latent semantic relations holding among context features. As shown in figure 1, this DSM will be employed at steps 1c and 3a for the computation of distributional representations for target sentences.

Step 1c: Computation of vectors for seed sentences. To assign a distributional representation to the seed sentences identified in step 1a, we adopt the strategy outlined in Schütze (1998). First, we turn sentences into bags-of-context-cues: e.g., sentence E_1 is assigned the set {*happen*, *night*}. The bag-of-context-cues is then turned into a sentence vector by averaging the distributional vectors of the words it contains.

Once the training material has been collected from the corpus and assigned a distributional representation, we proceed to the machine learning step:

Step 2: Training of an event/entity classifier with logistic regression. The sentence vectors computed at 1c are employed as learning events for a supervised logistic regression classifier, i.e. a machine learning tool which takes a set of distributional vectors and

the corresponding semantic classes as an input (in the example in figure 1, the pairs {$\vec{E}_1$, EVENT} and {$\vec{A}_1$, ARTIFACT}) and learns to weigh these features against each other to determine the probability of either class. In figure 1, the classifier learnt that *d*1 is highly indicative of an EVENT, that *d*2 is highly indicative of an ARTIFACT, and that *d*3 is just not discriminative enough.

In the evaluation step, the classifier is applied to the sentences contained in the dataset described above and evaluated against the manual annotation as follows:

Step 3a: Sentence vectors are computed for each sentence in the *-ment* dataset, according to the procedure described in 1c.

Step 3b: The classifier (step 2) is applied to the *-ment* sentence vectors. The outcome of this analysis is a probability distribution over the target semantic types for each instance of a derived word (e.g., 80% EVENT vs. 20% ARTIFACT). More balanced probability distributions arise from either uninformative or contradictory contexts.

Results. The experiments reported in this abstract are based on a "knowledge poor" approach to the task: when training our classifier (and, as a consequence, when classifying the target sentences) we did not resort to any linguistic pre-processing tools besides tokenization. The classifier was trained with logistic regression. We evaluated the logistic regression classifier with ten-fold cross-validation, turning the continuous predictions (from 0 to 1, with 1 corresponding to the event reading) into categorical responses by applying a threshold of 0.5. The classifier achieved a 73% accuracy (baseline=50). The very close accuracies of our classifier on the training and test sets indicate that it is not overfitting the data. We then applied the classifier to the sentences from the evaluation dataset. In the gold standard annotation, 250 instances are classified as events, 88 as objects, and 59 as ambiguous (10 sentences had to be discarded). In a binary classification task (event vs. object, threshold=0.5, majority class accuracy=74%) the classifier achieved 66.3% accuracy, with 19.2% false events and 15.5% false objects. In this setting, the ambiguous nouns have been mostly classified as events (66%). Further experiments involving the manipulation of the threshold did not outperform the majority class accuracy. Our interpretation is that the current model does not generalize well from the unambiguous training data to the test data (which are ambiguous, and therefore appear to employ different distributional contexts). In future work, we are conducting experiments in which training is focused on low frequency contexts in order to improve generalization to the test set.

Buitelaar, P. (1998). CoreLex: Systematic Polysemy and Underspecification. PhD Thesis, Computer Science Brandeis University.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis (special volume of the Philological Society), 1952-59, 1-32.

Kawaletz, L., & Plag, I. (2015). Predicting the semantics of English nominalizations: A frame-based analysis of -ment suffixation. In L. Bauer, L. Körtvélyessy, & P. Štekauer (Eds.), Semantics of complex words, 289-319. Dordrecht: Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR.

Lieber, R. (in press). English nouns: The ecology of nominalization. Cambridge: Cambridge University Press.

Plag, I., Andreou, M., & Kawaletz, L. (to appear). A frame-semantic approach to polysemy in affixation. In The lexeme in descriptive and theoretical morphology.

Schütze, H. (1998). Automatic word sense discrimination. Computational Linguistics 27(1), 97-123.