Concept-based Selectional Preferences and Distributional Representations from Wikipedia Articles

Alex Judea*, Vivi Nastase⋆, Michael Strube⋆

*University of Stuttgart
Stuttgart, Germany
alexander.judea@ims.uni-stuttgart.de

⋆ HITS gGmbH
Heidelberg, Germany
{vivi.nastase, michael.strube}@h-its.org

Abstract

This paper describes the derivation of distributional semantic representations for open class words relative to a concept inventory, and of concepts relative to open class words through grammatical relations extracted from Wikipedia articles. The concept inventory comes from WikiNet, a large-scale concept network derived from Wikipedia. The distinctive feature of these representations are their relation to a concept network, through which we can compute selectional preferences of open-class words relative to general concepts. The resource thus derived provides a meaning representation that complements the relational representation captured in the concept network. It covers English open-class words, but the concept base is language independent. The resource can be extended to other languages, with the use of language specific dependency parsers. Good results in metonymy resolution show the resource's potential use for NLP applications.

Keywords: dependency relations; selectional preferences; ontological and distributional meaning representations

1. Introduction

The NLP community has long explored various methods for representing word meaning. We capture word meaning in dictionaries, ontologies, or as points in multi-dimensional spaces, whose dimensions represent various facets of the word's behaviour in texts, based on Firth's observation that we can "know a word by the company it keeps" (Firth, 1957). The workshop series on Geometrical Models for Natural Language Semantics (GEMS)[1] show the continuing interest of the community in this issue.

We present such a distributional representation of concepts based on dependency relations obtained from Wikipedia articles. Concepts – roughly corresponding to Wikipedia articles – are represented through the open-class words with which they are grammatically related. For open-class words we compute their strength of association to general concepts, determining thus their selectional preferences. Resnik (1996) describes selectional preferences (also called selectional constraints or selectional restrictions) as limitations on the applicability of predicates to arguments. For example, the predicate expressed by the verb 'say' imposes constraints on the semantic classes of, e.g., its subjects: In the most frequent *literal usage* the required subjects of 'say' belong to the semantic class PERSON[2].

Our purpose for building a distributional representation connected to an ontology was to bring together two complementary meaning representations: while ontologies convey the meaning of concepts in terms of semantic relations to other concepts, grammatical collocations express the meaning of the concepts in terms of their involvement in events, or their functionality (e.g. an ARTICLE can be *written, published, read...*), and their attributes (it can be a *newspaper, journal, magazine* or *recent, excellent, short, prohibited ...* ARTICLE). And selectional preferences show the affinity of context for specific concepts.

Distributive representations of meaning assembled from collocations found in texts are subjected to one main critique – that they mix different senses for the target words. A partial solution for this comes from using topic models, where it is commonly assumed that the same word has different senses in different topics (Steyvers and Griffiths, 2007). By using Wikipedia and its "concept annotated" articles, as we will show in section 3. we bypass the ambiguity problem, at least with respect to the arguments of open-class words.

Distributive representations of meaning and selectional preferences can help in a variety of NLP tasks like word sense disambiguation, semantic role labeling, metonymy identification and resolution. We test the representation derived here, in particular the computed selectional preferences in the task of metonymy resolution, where we must determine whether a term has a literal interpretation – i.e. the meaning that fits the context best can be retrieved from a dictionary/sense inventory, or it has a particular figurative interpretation. To clarify, let us consider the sentence *I rode my Kawasaki on Sunday. Kawasaki* in this context refers to a product made by the Kawasaki company, as opposed to the company itself. Selectional preferences can be used to rank the possible interpretations of a word in context – in this case *rode/ride* prefers a vehicle as its object, rather than a company.

[1]The latest: `http://sites.google.com/site/geometricalmodels/`

[2]Expressions in SMALL CAPS denote concepts.

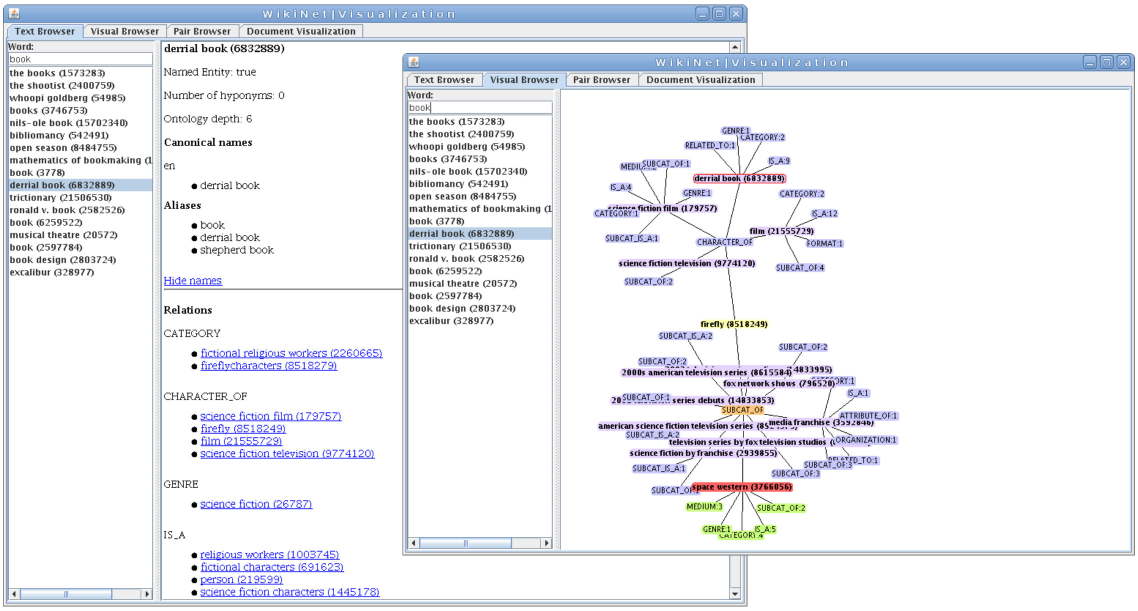

Figure 1: An example from WikiNet: snapshot of the word *book* and the chosen concept DERRIAL BOOK and its relations, partly expanded.

The paper will present first a brief overview of Wikipedia and WikiNet, the inventory of concepts used and to which the distributional representations derived are linked (Section 2.). We then describe how grammatical collocations are extracted and processed (Section 3.) to finally abstract concepts and compute selectional preferences for verbs and adjectives by generalizing their concept collocates using WikiNet (Section 4.). The metonymy resolution experiments will show the benefits of using the derived resource for the task (Section 5.).

2. Wikipedia and WikiNet

To build the representations and compute selectional preferences we use the set of English Wikipedia articles, which describe specific concepts. Wikipedia contributors are encouraged to insert hyperlinks, which link important terms in an article to the corresponding articles. A hyperlink consists of two parts, the actual link (i.e. a URL) and a phrase to appear in the text. Hyperlinks then constitute a bridge from the textual level to the conceptual level without the need for word sense disambiguation. We exploit these links in building the representations described in this paper.

The version we used (January 2011) provided over 3.5 million English articles, interconnected through a hierarchy of categories and hyperlinks. Based on these, Nastase et al. (2010) constructed WikiNet, a large-scale, multilingual semantic network whose nodes are concepts corresponding to articles or categories in Wikipedia. Concepts in this network are connected through a variety of semantic relations (e.g. *is a, member of, nationality*) derived from category names and infoboxes. WikiNet had 3,707,718 nodes and 49,931,266 relation instances of 494 types. Figure 1 shows a snapshot of a small part of the network produced by WikiNetTK, our toolkit for working with WikiNet[3].

We use WikiNet as a concept inventory, and its links and structure to generalize more specific concepts identified in texts to general concepts[4].

3. Extracting Concept Collocates

To compile our resource we tackle two sub-tasks: (1) extract predicate-argument collocations from a corpus and (2) abstract (generalize) the arguments extracted in step 1.

To extract grammatical collocations, all triples (w, r, c) or (c, r, w) are extracted from Wikipedia articles – w is an open-class word, c is a concept, and r is the grammatical dependency between w and c, e.g. (ride, *obj*, KAWASAKI).

We extracted the texts of all articles from an English Wikipedia dump[5], wikified it and split it into sentences. Wikifying an article is essentially linking phrases in the text to articles in Wikipedia (Mihalcea and Csomai, 2007). For this process we use the hyperlinks (anchor texts and targets) explicitly marked in the article, and apply the "one sense per discourse" assumption (Gale et al., 1992), whereby all occurrences of the hyperlink's anchor text in the same article refer to the same article/concept. Each expression found that refers to a concept is replaced with an expression that contains the concept's numeric ID (in WikiNet): e.g. CON_12 represents the concept with WikiNet ID 12 – ANARCHISM.

We collect all sentences that contain at least one concept (approx. 18 million), which are tokenized and POS tagged.

Tokenization and tagging was performed separately, as the

[3]WikiNet and WikiNetTK are available for download from `http://www.h-its.org/english/research/nlp/download/wikinet.php`.

[4]A tool kit for WikiNet is presented in Judea et al. (2011)

[5]The 2011/01/15 version.

Token	word	lemma	POS	POS	-	head ID	syn. role
1	CON_12	Anarchism	NN	NN	-	2	nsubj
2	is	be	VBZ	VBZ	-	0	root
3	often	often	RB	RB	-	2	advmod
4			...				

Table 1: Processing output in CoNLL format for the beginning of the sentence "Anarchism is often considered to be a radical egoism". 'CON_12' is the output of the wikification process and represents the WikiNet concept with ID 12, namely ANARCHISM.

parser used did not have embedded the corresponding components. This turned out to be an advantage, as we could control the assignment of part-of-speech tags to the concept ID – for concepts the tagger was presented with the corresponding concept name – which otherwise may have caused parsing problems. The results of preprocessing are represented in the EMNLP-CoNLL 2007 shared task format, as presented in Table 1.

The final step for collecting co-occurrence triples was parsing (columns 7 (head ID) and 8 (syntactic role) in the CoNLL format). Considering the amount of data (18 million sentences) it was crucial to select a fast (and reliable) parser. We use Ensemble (Surdeanu and Manning, 2010), which we parallelised to speed up the processing. 21 threads were used for parsing, 7 for Ensemble instances, each of which used 3 threads for the different parser models. In total, 150 GB of RAM were needed. The processing time was about 72 hours, parsing about 70 sentences per second on average.

Figure 2 shows an example of grammatical collocations for the verb *say* obtained after the processing described above.

say
nsubj
Rick Martin (3457416), Roosevelt (10979), Jesus (1095706), G.I. Joe (28713688), Teletubbies (31015), Portugal (23033), ...
prep_to
Lizo Mzimba (2113568),
Ralph Allen (1121430), ...,
Minerva (19845), ...,
Mauritius (19201), ...

Figure 2: Concept collocates for the verb "say" and the dependencies *nsubj* and *prep_to*. Numbers in parentheses are WikiNet IDs.

4. Abstraction for Selectional Preference Computations

The extracted collocations can be used to represent the concepts in the data, and can be further processed to compute selectional preferences. For this, the extracted concept collocates were abstracted to more general concepts: e.g. for the verb *say* (in Figure 2) Rick Martin, Roosevelt, and Jesus are all person, which represents better the selectional preference in the subject position for this verb than the individual persons.

The question is how to determine the most appropriate abstraction level. For example, Rick Martin could be abstracted to hockey players, person, or concept[6]. Intuitively, for 'say' and the role *nsubj*, Rick Martin should rather be abstracted to person than to hockey players or concept.

To find an appropriate trade-off between abstraction and expressiveness, we developed the following three-step algorithm, performed for each open-class w and grammatical relation r:

1. gather a list of candidate abstractions, $\mathcal{A}_{w,r}$,
2. prune the list,
3. compute a score for each abstraction based on a measure of semantic relatedness.

Steps (1) and (2) determine a list of abstractions, step (3) outputs the scores of abstractions given a specific collocate.

For step (1), given an open-class word w and a grammatical dependency r, each collocated concept (c, r, w)[7] is expanded based on its *is_a* relations in WikiNet, up to a maximum number of steps (from our previous experience with WikiNet, we set this number to 4). The result is a list of abstraction candidates with associated frequencies.

In step (2) the list is pruned according to the score s_h of each abstraction candidate h, which is measured in terms of ontology depth, number of hyponyms, and relative number of votes of h:

$$s_h = \frac{1}{depth(h)} + \frac{\log(hyponyms(h)+1)}{\log(|concepts(WN)|)} + \frac{votes(h)}{|votes|} \quad (1)$$

$depth(h)$ denotes the estimated depth of h computed based on the category hierarchy. The second term expresses the generality of the candidate as a function of the number of its hyponyms in WikiNet. The third term is the relative amount of votes the candidate received. Pruning is done by filtering out concepts with $s_h < \tau$. If there are no concepts with a score above τ, the threshold is gradually decreased until some abstractions pass it or $\tau = 0$. The candidates that passed the threshold constitute the set of abstractions $\mathcal{A}_{w,r}$.

The final step involves scoring the selected abstractions using a measure of semantic relatedness based on concept paths in WikiNet. It aims to assign a higher score to those abstractions which are more strongly related to the collocates they stem from. The number of paths from a concept collocate to an abstraction is a good indicator of relatedness, though longer paths are weaker indicators than shorter ones. The path score – ps – formalizes these propositions:

[6]In WikiNet, person *is_a* concept.

[7](w, r, c) triples are treated in a similar manner.

Algorithm 1 Determine abstractions (general concepts) for an open-class word w and a grammatical role r, and compute selectional preferences with respect to w and r.

Input:
$\mathcal{W}_{cooc} = \{(c, r, w) or (w, r, c) \in Wikipedia,$
c is a concept, w is an open class word,
r is a grammatical relation$\}$,
$WikiNet$ – a concept network extracted from Wikipedia, with is_a relations
Output: $\mathcal{A}_s$ – a set of scored abstractions

$\mathcal{A}_s = \{\}$
for all (w, r) such that $(*, r, w) \in \mathcal{W}_{cooc}$ **do**
$\mathcal{C}_{w,r} = \{c | (c, r, w) \in \mathcal{W}_{cooc}\}$
$\mathcal{A}_{w,r} = gatherAbstractions(\mathcal{C}_{w,r}, WikiNet)$;
$\mathcal{A}_{s/w,r} = scoreAbstractions(\mathcal{A}_{w,r}, WikiNet)$;
$\mathcal{A}_s = \mathcal{A}_s \cup (w, r, \mathcal{A}_{s/w,r})$
return $\mathcal{A}_s$

gatherAbstractions

Input:
$\mathcal{C}$ – a set of concepts,
$WikiNet$ – a concept network with is_a relations
Output: $\mathcal{A}$ – a set of abstractions

for all $c \in \mathcal{C}$ **do**
for all hypernyms h : (c, is_a, h) $\in$ WikiNet **do**
$\mathcal{A} \cup (h, votes(h) + 1)$
while $\mathcal{A} = \emptyset \wedge \tau \geq 0$ **do**
for all (h, votes(h)) $\in \mathcal{A}$ **do**
$s_h = \frac{1}{depth(h)} + \frac{\log(hyponyms(h)+1)}{\log(\#\ concepts)} + \frac{votes(h)}{\#\ votes}$
if $s_h \geq \tau$ **then**
$\mathcal{A} \cup (c, s_h)$
if $\mathcal{A} = \emptyset$ **then**
$\tau = \tau - 0.1$
return $\mathcal{A}$

scoreAbstractions

Input:
$\mathcal{A}$ – a set of abstractions
$WikiNet$ – a concept network with is_a relations
Output: $\mathcal{A}_s$ – a set of scored abstractions

for all abstractions $a \in \mathcal{A}$ **do**
$\mathcal{P}^a_c$:= WikiNet paths between a and all $c \in \mathcal{W}_{cooc}$
$\mathrm{ps}_a = 0$
for $l = 1 \ldots max\_path\_length(\mathcal{P}^a_c)$ **do**
$pl := number\ of\ paths\ with\ length\ l$
$\mathrm{ps}_a = \mathrm{ps}_a + (\frac{pl}{l})$
$\mathcal{A}_s \cup (a, \mathrm{ps}_a)$
return $\mathcal{A}_s$

$$\mathrm{ps}_{a,c} = \sum_{i=1}^{max_i} \frac{|p_i|}{i} \quad (2)$$

with

$$\{p_i \subset \mathcal{P}^a_c | length(p) = i\}. \quad (3)$$

$\mathcal{P}^a_c$ is the set of paths from the concept c to an abstraction a, p_i is the subset of paths of length i, and max_i is the length of the longest path in $\mathcal{P}^a_c$.

After $ps_{a,c}$ is computed for every $a \in \mathcal{A}$ and concept c in relation (c, r, w) for a given word-relation pair (w, r), the scores are aggregated over all c, resulting in a scored list illustrated in Figure 3.

say
nsubj
PERSON 29.777
GEOGRAPHY 4.108
SOCIETY 4.021
...
prep_to
PERSON 13.476
...
announce
nsubj
PERSON 48.243
ORGANIZATION 6.632
POLITICS 6.418
...
prep_in
SOCIETY 6.050
GOVERNMENT 5.726
...
beautiful
amod
CULTURE 2.396
PERSON 2.37
...
big
amod
GEOGRAPHY 2.447
CULTURE 1.921
...

Figure 3: Examples of scored abstractions as selectional preferences.

Selectional preferences for adjective arguments are weaker than those for verbs, indicating that the same adjective can modify a more varied array of concepts than verbs do.

The method presented here can distinguish between competing abstractions. For example, it determines that PERSON is preferred to CONCEPT and LEADERS, given *say* and the dependency *nsubj*. It can also abstract *single* concepts (as opposed to pairs or sequences of concepts). Algorithm 1 formalizes this.

5. Evaluation through metonymy resolution

The resource built (cSR) was evaluated as part of a metonymy resolution system. Metonymy is a pervasive phenomenon in natural language, whereby a term is used to express a (semantically) related concept. In most cases of metonymy, the relation between the metonymic term and what it refers to is systematic: a company name is used to refer to its products, location, employees, etc.

The selectional preferences were used to compare, for a potentially metonymic word (PMW), its literal interpretation and the preferred meaning based on its grammatically related verbs and adjectives which induce the *required reading* of the PMW.

Consider the example: `I bought a new BMW`. Here, *BMW* is metonymic, because not the BMW company is

XML tagged text

<sample id="samp114">
<bnc:title> Computergram international </bnc:title>
<par>
LITTLE FEAR OF MICHELANGELO
The computer industry equivalent of "Small earthquake in Chile" ...
The Michelangelo computer virus that received worldwide attention last year is expected to cause even fewer problems this Saturday than it did when it struck last year, a team of <annot><org reading="literal"> IBM </org></annot> researchers said.
</par>
</sample>

Figure 4: Sample annotation

meant, but a car produced by it. The noun is syntactically linked to the verb *bought* through a direct object relation and to the adjective *new* through an adjectival modification. The highest-scored restriction of *bought/buy* in cSR is ORGANIZATION, the highest-scored restriction of *new* is PERSON, but the latter does not belong to both *buy* and *new* in the resource. What the two have in common are the restrictions, e.g., ORGANIZATION and CULTURE. The individual scores they have for the restrictions are summed and CULTURE is selected as the highest-scored restriction – this is considered to be the required reading of *BMW*. Note that in most cases the computed required reading is not an appropriate literal interpretation of the potentially metonymic word but lies on a more general level in the semantic network.

Based on required readings (and other sources of evidence, like the syntactic context of a PMW), a metonymy prediction model is trained. We work with the data from the metonymy resolution task at SemEval 2007 (Markert and Nissim, 2007), generated based on a scheme developed by Markert and Nissim (2003).

The metonymy resolution task at SemEval 2007 consisted of two subtasks – one for resolving country names, the other for companies. For each subtask there is a training and a test portion. Figure 4 shows the text fragment for one sample, and Table 2 the data statistics. The *reading* column shows the possible interpretations of a PMW for countries and companies respectively. For example, `org-for-product` would be the interpretation of the PMW *BMW* in the previously shown example. The task was three-fold: In the 'coarse' task, only literality or non-literality was to be predicted, in the 'medium' task literality was to be distinguished from the categories 'metonymy' and 'mixed' (if a PMW has a literal interpretation according to part of the context, but metonymic according to the rest), finally, the 'fine' task required systems to resolve the broad category 'metonymy' to fine-grained metonymy classes like organization-for-product (org-for-product).

To test the usefulness of the selectional preferences computed, we add them to a set of features commonly used for the metonymy resolution task, proposed first by Nissim and Markert (2005):

reading	train	test
locations	925	908
literal	737	721
mixed	15	20
othermet	9	11
obj-for-name	0	4
obj-for-representation	0	0
place-for-people	161	141
place-for-event	3	10
place-for-product	0	1
organizations	1090	842
literal	690	520
mixed	59	60
othermet	14	8
obj-for-name	8	6
obj-for-representation	1	0
org-for-members	220	161
org-for-event	2	1
org-for-product	74	67
org-for-facility	15	16
org-for-index	7	3

Table 2: Reading distributions

- grammatical role of PMW (subj, obj, ...);
- lemmatized head/modifier of PMW (announce, say, ...);
- determiner of PMW (def, indef, bare, demonst, other, ...);
- grammatical number of PMW (sg, pl);
- number of grammatical roles in which the PMW appears in its current context;
- number of words in PMW;

We add a feature for each abstraction a associated with each of the open class word w through the same grammatical relation r with which w and the annotated PMW are related in the data. For each PMW, its grammatically related open-class w and each abstraction a, the value of the corresponding feature a is the computed selectional preference from w to a.

Table 3 presents accuracies for the medium granularity for several configurations: *bl.* – baseline: all test instances are assigned the most frequent category 'literal'; *SemEval* – the best result obtained by a system in task8 at SemEval 2007; *M&N* – results obtained using only the features proposed by Nissim and Markert (2005); *cSR* – the metonymy resolution system when using only the derived abstractions as features (and their selectional preferences as feature values); *cSR+M&N* – the system when it combines abstractions and selectional preferences, with the M&N features.

domain	bl.	SemEval	M&N	cSR	cSR+M&N
org.	61.8	73.3	69.4	69.1	72.0
location	79.4	84.8	82.3	82.3	85.6

Table 3: Accuracies for the medium-grained metonymy resolution task in two domains, organizations and locations. The baseline is the most frequent class in the test data ('literal').

Using selectional preferences clearly improves accuracy results over the baseline, and in combination with the frequently used contextual features, it improves over the best result obtained in the metonymy resolution task at SemEval 2007 for the location domain, but lags behind on the organization domain. One of the reasons for these results is the fact that our process relies on automatically derived resources (both WikiNet and grammatical collocations from Wikipedia articles were automatically extracted), which are not perfect. We plan to further explore the task and investigate the reason for the difference in performance on the two domains, and develop more robust methods that can deal with noisiness in the grammatical collocations and hierarchical relations in WikiNet.

6. Discussion and conclusions

The resource described here[8] provides semantic descriptions for concepts in terms of their syntactic collocations, and concept-based selectional preferences for open-class words.

The distributional semantic representation thus obtained complements an ontology (concept network) obtained from processing various facets of Wikipedia – category-article network, infoboxes, disambiguation, redirect and cross-language links, etc. The combination of the two meaning representations have allowed us to obtain selectional preferences from collocations, by abstracting collocation instances based on a concept's position and relations in the concept network. The resource obtained was evaluated through metonymy resolution, and the results indicate that the computed preferences are reliable and useful.

We plan to extend the resource by extracting and computing preferences within verb frames, instead of single relations, to expand it to other languages as competitive taggers and parsers become available, and to try other, probability-based, scoring methods for abstractions.

Acknowledgments

This work has been financially supported by the EC-funded project CoSyne (FP7-ICT-4-24853) and the Klaus Tschirra Foundation. We also thank the organizers of the SemEval-2007 Task #8 *Metonymy Resolution* – Katja Markert and Malvina Nissim – for making the annotated data freely available.

And finally, we are ever grateful to the WikiMedia organization for setting up the collaboration platform and Wikipedia's anonymous contributors who, inadvertently, have provided the NLP community with a treasure trove of knowledge.

[8]Available for download from `http://www.h-its.org/english/research/nlp/download/wikinet.php#cSR`.

7. References

John Rupert Firth. 1957. Modes of meaning. In *Papers in Linguistics 1934-1951*, pages 190–215. London, U.K.: Longmans.

Willian A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

Alex Judea, Vivi Nastase, and Michael Strube. 2011. WikiNetTk – A tool kit for embedding world knowledge in NLP applications. In *Proceedings of the IJCNLP 2011 System Demonstrations,* Chiang Mai, Thailand, 9 November 2011, pages 1–4.

Katja Markert and Malvina Nissim. 2003. Corpus-based metonymy analysis. *Metaphor and Symbol*, 18(3):175–188.

Katja Markert and Malvina Nissim. 2007. SemEval-2007 Task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1),* Prague, Czech Republic, 23–24 June 2007, pages 36–41.

Rada Mihalcea and Andras Csomai. 2007. Linking documents to encyclopedic knowledge. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM 2007),* Lisbon, Portugal, 6–9 November 2007, pages 233–242.

Vivi Nastase, Michael Strube, Benjamin Börschinger, Cäcilia Zirn, and Anas Elghafari. 2010. WikiNet: A very large scale multi-lingual concept network. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* La Valetta, Malta, 17–23 May 2010.

Malvina Nissim and Katja Markert. 2005. Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy. In *Proceedings of the 6th International Workshop on Computational Semantics,* Tilburg, Netherlands, January 12-14, 2005.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, (61):127–159, November.

Mark Steyvers and Thomas Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, Hillsdale, NJ.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble Models for Dependency Parsing: Cheap and Good? In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 2–4 June 2010, pages 649–652.