Enriching WordNet Via Generative Metonymy and Creative Polysemy

Jer Hayes*, **Tony Veale***, **Nuno Seco***

*Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
{jer.hayes, tony.veale, nuno.seco}@ucd.ie

Abstract

Metonymy is a creative process that establishes relationships based on contiguity or semantic relatedness between concepts. We outline a mechanism for deriving new concepts from WordNet using metonymy. We argue that by exploiting polysemy in WordNet we can take advantage of the metonymic relations between concepts. The focus of our metonymy generation work has been the creation of noun-noun compounds that do not already exist in WordNet and which can be profitably added to WordNet. The mechanism of metonymy generation we outline takes a source compound and creates new compounds by exploiting the polysemy associated with hyponyms of the head of the source compound. We argue that metonymy generation is a sound basis for concept creation as the newly created compounds are semantically related to the source concept. We demonstrate that metonymy generation based on polysemy is superior to a method of metonymy generation that ignores polysemy. These new concepts can be used to augment WordNet.

1. Introduction

One of the advantages of WordNet (WN) (Miller, 1995) is that it is a rich lexical knowledge base. However, in practical terms a taxonomic lexical database will be by its nature incomplete. This is due to a number of factors: time constraints, financial constraints and that language is a dynamic system (new words and senses frequently enter the language). Ontologies, in general are time-consuming to create and may be financially burdensome to develop and support. WN is an ontology that represents a synchronic snapshot of the English lexicon, an inherently diachronic system. As such, large ontologies of the English lexicon can only hope to capture a salient selection of the concepts that should be represented. For WN, the selection criteria are determined by conventional word usage, but many concepts that can profitably be represented are omitted. These omissions lead to holes and asymmetries in the ontology that can significantly mislead automated reasoning systems that are sensitive to the organization of the ontology.

WN makes no distinction between homonymy and polysemy. For example, the orthographical word bank has senses associated with it that are related to a financial institution and a land structure. This would not be the case in traditional dictionaries where the orthographical word bank when treated as noun may be associated with three different noun entries (e.g. see www.m-w.com) and for each noun entry polysemous senses are grouped together. By classifying which senses are related by polysemy we can create metonymies that will be useful in many NLP applications and esp. in query expansion. Polysemy is an untapped resource within WN and so in this paper we outline how information on polysemy can be extracted and we outline one of the possible applications of this information, metonymy generation. The basis of our approach lies in the insight Apresjan had into systematic polysemy (Apresjan, 1974), namely that the regularity of polysemy in the lexicon can be considered the result of regular metonymy processes. The regularity of polysemy is displayed in the identifiable families of words that each exhibit similar sense patterns. Consequently and crucially, this means that by examining systematic polysemy we can examine metonymy.

Metonymy has been defined as the using of one entity to refer to another that is related to it (Lakoff and Johnson, 1980) but more specifically we can say an expression A is a metonym, if A deviates from its literal denotation in that it stands for an entity B, which is not expressed explicitly but is related to A via a metonymic relation (Markert and Hahn, 1997). Metonymy is also one of the major processes of systematic sense extension (Sweetser, 1991) (along with specialisation and generalisation) and so is often a direct cause of polysemy. Consequently and crucially, by examining systematic polysemy we can gain insight into metonymy. For example, taking a word such as *music* from WN we can ascertain that *dance* is a metonymically related word, as is *composer*. These metonyms can be found by examining the concepts that are types of music and examining what other categories these concepts fall into. Further to this we suggest that the lexico-conceptual information that can be drawn from polysemy points to possible metonymic links between concepts. For example, a *composer* can be a metonym for her *music* and *music* can be a metonym for the *dance* carried out to that type of music.

The mechanism of metonymy generation we outline creates compounds such as “farm device” which do not exist in WN. Metonymy generation involves taking a source concept, e.g. “farm worker” in WN and finding candidate metonymies for this source concept, e.g. “vehicle”, as some types of worker also name associated devices. For example the word *peeler* has three senses in WN two of which are: “a worker who peels the skins from fruits and vegetables” and “a device for peeling vegetables or fruits”. Where senses are linked via polysemy, i.e. the sense are related then a compound is created via the head of the related sense. These candidate metonymies are then validated in relation to WN or to web-based documents. As metonymy is based on contiguity the validated concepts should be semantically related to the source concept. The generation of metonymies in everyday language is a creative process; by employing metonymy generation we can apply a potent creative process as a process of concept creation and we can find ontological holes in WN.

1.1. Goal of the paper

In this paper we will outline how metonymy generation operates when the source concept is a literal compound and the metonymies generated are also compounds. We also examine two strategies for metonymy extraction one which relies on polysemy and on which does not. Using two different methods of validation we demonstrate that the metonymy generation process based on polysemy is more precise than one which ignores polysemy.

2. Exploring Polysemy in WordNet

The principle behind systematic polysemy is that the nature of the relationship between the senses of one word may hold for many. The original definition of systematic polysemy or regular polysemy was given by Apresjan (Apresjan, 1974): "Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and and if a_i and b_i, a_j and b_j are nonsynonymous", (p16). In WN both the words *Bantu* and *Algonquian* have two senses and each sense for both words refers to a people and a language. According to the definition of Apresjan above these words display regular or systematic polysemy. For Apresjan regularised sense extension was largely a question of metonymic transfer.

Previous attempts at exploring polysemy within WN have relied on the structural patterns that polysemous senses exhibit within WN (Peters, 2002; Peters and Peters, 2000). The patterns investigated generally involved the search for concepts which had senses that fell into two separate parts of the WN taxonomy and checking if sibling concepts also fall into similar categories. However, many other structural patterns exist within WN. For example many polysemous words in WN exhibit the following pattern: one sense, S1, names a parent of another sense, S2, in its gloss. This is the case with the word *Uighur* in Table 1. The word *Uighur* has senses that belong to different parts of the WN taxonomy, language and people. In addition, one of the glosses for *Uighur* mentions a parent of another sense. We have discovered several of these polysemy patterns but we will focus on the pattern we have dubbed "cross-referencing" (Veale, 2003).

3. Generating Compounds via Metonymy

New compounds are found by analyzing the descendents of the head of a literal compound. A literal compound is the immediate child of its head. Where the descendents of a literal compound have more than one sense and these additional senses are not also descendents of the head then the parents of this child are used to form a compound with the modifier of the original compound. Ideally we wish to find parents of senses that are polysemous. Yet we can also generate compounds without using polysemy. Compounds generated via polysemy should be more effective in finding new compounds, and so will be more likely to validated, than those that are generated by a process not guided by polysemy and we investigate this proposal in this paper.
There are two broad strategies we will examine in relation to the generation of metonymic compounds, one based on polysemy and one which treats all ambiguities as possibly polysemous -

Ambiguity Strategy: For every compound in WN of the form "M_H" and where M and H are entries in WN, if "M_H" has a hypernym which is a sense of H, H_i*, then find every descendent,* D_k*, of* H_i*. If* D_k *has two senses with meanings [*D_x*,*D_y*], then for every alternate sense* D_k *consider every parent hypernym* P_k *of* D_k*. If* P_k *is not also a hypernym of* D_j *(the descendent of* H_i*) And if the word P is the head of some WN compound X-P generate the hypothetical concept M-P.*

Polysemy Strategy: For every compound in WN of the form "M_H", where M and H are entries in WN. If "M_H" has a hypernym which is a sense of H, H_i*, then find every descendent,* D_k*, of* H_i*. If* D_k *has two senses with meanings [*D_x*,*D_y*] and If and only If these senses are polysesmous Then for every alternate sense* D_k *consider every parent hypernym* P_k *of* D_k*. If Pk is not also a hypernym of* D_j *(the descendent of* H_i*) and if the word P is the head of some WN compound X-P generate the hypothetical concept M-P.*

Given the literal compound "medical specialist", the algorithm for strategy 1 would generate some of the following compounds: "medical historian", "medical expert", "medical analyst", "medical host", "medical architect", "medical painter". Given the same compound, polysemy would produce: "medical designer". This suggests that strategy 1 will produce more new compounds but many of these may not be related to the original compound as they ignore polysemy.

4. Validating new compounds

We adopt two mechanisms for validating a new compound: (1) internal validation - where a compound points to an existing concept in WN, and (2) external validation - where a compound exists in a number of web documents above a specified threshold. Internal validation is found as follows: given a compound of the form "M-H", if M is listed in a gloss of one of the descendents of H then it points to this concept, e.g., the compound "farm business" would point to the concepts ["animal husbandry", "mixed farming"], as all are types of business which list farm in their respective glosses. Note that this new compound may also be usefully added to WN as the category "farm business" does not exist in WN.

The web has been used a corpus for a number of traditional NLP tasks, e.g. example-based machine translation (Way and Gough, 2003), statistical-based translation (Kraaij and Simard, 2003) and likewise we use the web as a corpus for validating new compounds. Essentially, external validation uses web-based documents to ascertain if the new compound already exists. Given a new compound we submit it to the AltaVista search engine and record how many sites this new compound appears in. This submission looks for the exact phrase of the compound within docu-

Pattern			Example		
P1		P2	Language		People
⇕		⇕	⇕		⇕
S1	“..P2..”	S2	Uighur: S1	“..People...”	Uighur: S2

Table 1: “Cross-reference” polysemy pattern

ments.[1] Before we examine results of the comparison between polyemy strategy and ambiguity strategy let us suggest that if a compound exists it should be found on the web. For example, taking all the existing literal compounds in WN we can ascertain how many sites these concepts are listed on. There were a number of compounds found which did not match with any documents within the AltaVista index and these compounds were almost always zoological terms, e.g. “genus Amphicarpa” returned no results. However, the average number of documents matched was 37,850. This suggests that external validation will be useful in judging whether the new compound generated refers to something that exists or if it does not (and is thus, probably nonsensical or in extremely low usage).

In relation to concept creation and WN both validation procedures offer different advantages. Internal validation finds concepts that should immediately fit neatly into WN. External validation finds concepts which have not been included in WN either for purposes of economy of space or because they are new concepts. It may be more cumbersome to add these concepts to WN. However, both internal and external validation show up the ontological gaps in WN.

5. Results

Having devised two strategies for finding new compounds, one of which exploits polysemy, and outlining two mechanisms for validating these new compounds we now outline the testing of these two strategies. This testing was carried out in two phases. In phase I, the first strategy was tested in phase II the second strategy was tested. Each phase involved the generation of new compounds and the validation of all newly generated compounds both internally and externally. We will foucs our analysis of these strategies into on compound types created. As we suggest that compound types offers a fairer basis of comparison. The ambiguity strategy produces a large number of new compound tokens compared to compound types. For example, of the first 290,000 new compound tokens produced by the ambiguity strategy over 7,000 new compound types were found. The polysemy strategies in contrast have a closer ratio of new compound types to new compound tokens. We have chosen to focus on the results of validation on new compound types, although, this will mask the large amount of redundant information the ambiguity strategy will produce.

The ambiguity strategy creates the greatest number of new compound types, 459,772. But these new compound types are also the least likely to be externally validated when compared to the polysemy strategy. Of the 459,772 compounds 15,215 were externally validated. Thus, 42% of new compound types derived from this strategy could be externally validated. Only, 7,721 compound types were internally validated, so, only 11% of compound tpyes were internally validated.

The Polysemy strategy produces new compounds based on sense pairings from the cross-reference pattern and it generates 1,418 new compound types. This is far lower than the ambiguity strategy, however, these compounds are more likely to be externally validated than the ambiguity strategy. 74.4% of new compound types were externally validated. 1,056 new compounds were externally validated. This strategy also produced the highest number of internally validated compounds relative to the total number of new compounds produced. Of the initial 1,418 compound types 31% were internally validated. Overall, the polysemy strategy is more effective in producing compounds that exist outside the WN ontology than the ambiguity strategy. However, no strategy was entirely effective in finding compounds that do not exist in WN at present but which can be internally validated with respect to WN. The difference between the internal and external validation rates for each strategy suggest that there are a good deal of noun-noun compounds which exist outside of WN and which cannot be internally validated by WN. This is not surprising as WN is a general-purpose ontology and will have ontological holes.

New compounds which are validated internally can be directly fitted into WN whereas those externally validated will require more work to fit in. For example, the literal compound “farm worker” gives rise to some of the following internally validated concepts: “farm device”, “farm laborer”, “farm group”. These new compounds can act as hypernyms for the concepts which they appear to name. In this case “Farm device” would be a hypernym of the concepts: harvester, reaper, haymaker, “hay conditioner”, thresher, thrasher, “threshing machine”, cultivator, and tiller.

In general, externally validated concepts will not easily fit into WN, therefore we can question the utility of the externally validated compounds with respect to WN. To judge the utility of these new compounds we analysed those compounds created via the polysemy strategy and deduced whether the new compounds were related to the source literal compound. Overlall, we judged 447 source literal compounds with 1,972 new compounds (some new compounds

[1] This may result in some false positives but by setting a high threshold of 40 documents these should be removed.

	Ambiguity	Polysemy
Total number of new compound types found:	459,772	1,418
Total number of new compound types internally validated:	52,215	451
% of new compound types internally validated:	11%	31%
Total number of new compound types externally validated:	195,307	1,056
% of new compound types externally validated:	42%	74.4%

Table 2: Findings for phase I and phase II

were created more than once, hence the descrepancy with the figure of 1,418 in Table 2). Of these initial 447 source literal compounds, 379 literal compounds with 1,014 expansions were judged to be relevant. Some examples of these revelevant new compounds are:
“scrub_brush”: “scrub_worker”
“coffee_tree”: “coffee_seed”
“wedding_ceremony”: “wedding_ritual”, “wedding_rite”
“business_establishment”: “business_job”
Although these new compounds may not have a good coverage as hypernyms, these new compounds should still be associated with the original source word in WN, as they may prove useful for NLP tasks such as query expansion. This association could be implemented as a layer on top of WN and would not require a restructuring of the ontology.

6. Conclusions

We suggested that metonymy generation which exploits polysemy should be more effective than a strategy based on ambiguity in terms of how many compounds are validated. We proposed two validation techniques: (1) internal validation and (2) external validation. The polysemy strategy was more effective in creating compounds that were both internally and externally validated.

All knowledge-based systems by their nature will have ontological gaps. By using a mechanism such as metonymy generation we can create new concepts that can be added back into knowledge base. The concepts can act as new categories. For example, we suggested that “farm vehicle” can be a new type of vehicle which will serve to link types of vehicle which are related to the concept farm. However where a compound does not act as a category for concepts in WN its addition to WN may still be useful. For example, the new compound may be useful in terms of query expansion.

Our model of metonymy generation exploits the polysemy inherent in WN. We have discovered several patterns of polysemy within WN, although in this paper we have focused exclusively on just one of these. The metonymy generation mechanism that is based on polysemy has not as yet been tested on other polysemy patterns. It may be the case that some of these patterns produce more favorable results. So in future work these polysemy patterns should be tested against each other in terms of metonymy generation.

7. References

Apresjan, J., 1974. Regular polysemy. *Linguistics*, 142:5–32.

Kraaij, Nie J., W. and M. Simard, 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29(3):381–419.

Lakoff, G. and M. Johnson, 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Markert, K. and Udo Hahn, 1997. On the interaction of metonymies and anaphora. In Martha Pollack (ed.), *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann.

Miller, G., 1995. Wordnet: A lexical database of english. *Communications of the ACM*, 38:39–41.

Peters, W., 2002. Extraction of implicit knowledge from wordnet.

Peters, W. and I. Peters, 2000. Lexicalised systematic polysemy in wordnet.

Sweetser, E., 1991. *From etymology to pragmatics*. Cambridge: Cambridge University Press.

Veale, T., 2003. Pathways to creativity in lexical ontologies.

Way, A. and N. Gough, 2003. webmt: Developing and validating an example-based mt system using the world wide web. *Computational Linguistics*, 29(3):421–457.