Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded

Marius Paşca

Google Inc., 1600 Amphitheatre Parkway,
Mountain View, California 94043
mars@google.com

Abstract. This paper presents an approach to extending existing lexical resources with instance names and alternative definitions acquired from textual documents. The experiments involve WordNet and approximately 300 million Web documents, but the method is more generally applicable. We leverage formally-structured, human-validated resources, on one hand, and data-driven instance names and definitions on the other, which opens the path to new applications of the reloaded resources.

1 Motivation and Goals

Large-scale lexical, hierarchical resources have a broad range of applications in computational linguistics, information extraction and information retrieval. When manually building such resources, the focus is justifiably on selecting and organizing words into hierarchies of conceptual entries, with manual selection of an ideal, single definition for each entry. For example, by grouping together English words with the same meaning (e.g., *lawyer* and *attorney*) into sets of synonyms (or *synsets*, such as {*lawyer, attorney*}) associated with a single definition (or *gloss*), WordNet [1] became a de-facto standard for lexical resources. Its uses span word sense disambiguation [2], information extraction [3] and machine translation [4], to name only a few.

Hierarchical resources organize noun synsets along *IsA*/*InstanceOf* relations. The conceptual coverage of WordNet is impressive, with more than 150,000 English words encoded in over 115,000 synset entries or lexical concepts - more than half of which are nouns. However, WordNet and other resources are not necessarily complete for obvious practical reasons. This particularly applies to the lower-level hierarchies, where the more specific concepts occur, in the form of both missing specialized concepts and missing instance names. WordNet does not contain *telecom company* or *meta search engine* under *company* and *search engine* respectively; similarly, there are no instance names such as *Google* under *search engine*, or *Ferrari* under *car company*. Only a fraction of the encoded concepts are accompanied by corresponding instances; the number of such instances embedded under a given concept is usually small. For instance, 600 instance names exist under *city*; comparatively, there are eight instance names under *lawyer* (including *Francis Scott Key* and *Abraham Lincoln*), one instance

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 280–292, 2005.

name under *skyscraper* (*World Trade Center*), and one instance name under *cavern* (*Carlsbad Caverns*). The first goal of this paper is to expand lower-level hierarchies with instance names acquired from textual Web documents.

In most resources, including WordNet, the lexical concept entries contain single rather than multiple strings as definitions. For example, *machine translation* is defined in WordNet as *"the use of computers to translate from one language to another"*. Since definitions are unique per word sense, higher-level applications that operate on them, e.g. lexical chains [5] or semantic similarity measures [6], can rely only on the particular sequence of words actually included in the definitions. However, usually there is more than one way to express the same definition. As an illustration, alternative definitions for *machine translation"* include *"process of translating documents from one language to another by computer"*; *"process by which a machine translates text from one language to another"*; and *"automatic translation of human language by computers"*. [1] The alternative definitions will capture various morphological, lexical and semantic variations, and make them available to the higher-level applications. This also represents a novel source for extracting paraphrases, which are useful in information extraction, document retrieval and question answering. The second goal of this paper is to use Web textual documents to extract alternative glosses or definitions for existing concepts situated in lower-level noun hierarchies.

The remainder of the paper is structured as follows. After an overview of the method in Section 2, Section 3 describes the identification and extraction of relevant text nuggets from unstructured text. The nuggets are the raw material for deriving alternative glosses and new instance names, as shown in Section 4. Section 5 describes experiments on approximately 300 million Web documents. After further discussion in Section 6, we conclude in Section 7.

2 Method at a Glance

The method and experiments described in this paper augment the concepts situated in the lower-level WordNet hierarchies with new information, namely instance names and alternative glosses. As an illustration, the top box (A) in Figure 1 contains part of the original WordNet hierarchy under the lexical concept of *computer program*. In contrast, the other boxes (B, C and D) in the figure contain some of the new information actually extracted from Web documents.

The main source of the newly acquired information are *text nuggets*, which are sentence fragments extracted uniformly from unstructured text. To identify relevant nuggets across the Web, we focus on textual content rather than structural clues. Shallow lexico-syntactic extraction patterns are applied to the unstructured text of Web documents. The extraction patterns are designed to be lightweight and simple to handle robustly the noise and diversity of Web documents. The conversion of the extracted text nuggets into higher-level information, namely instance names and glosses, also relies on minimal text processing

[1] These definitions were extracted from the Web with the method described herein.

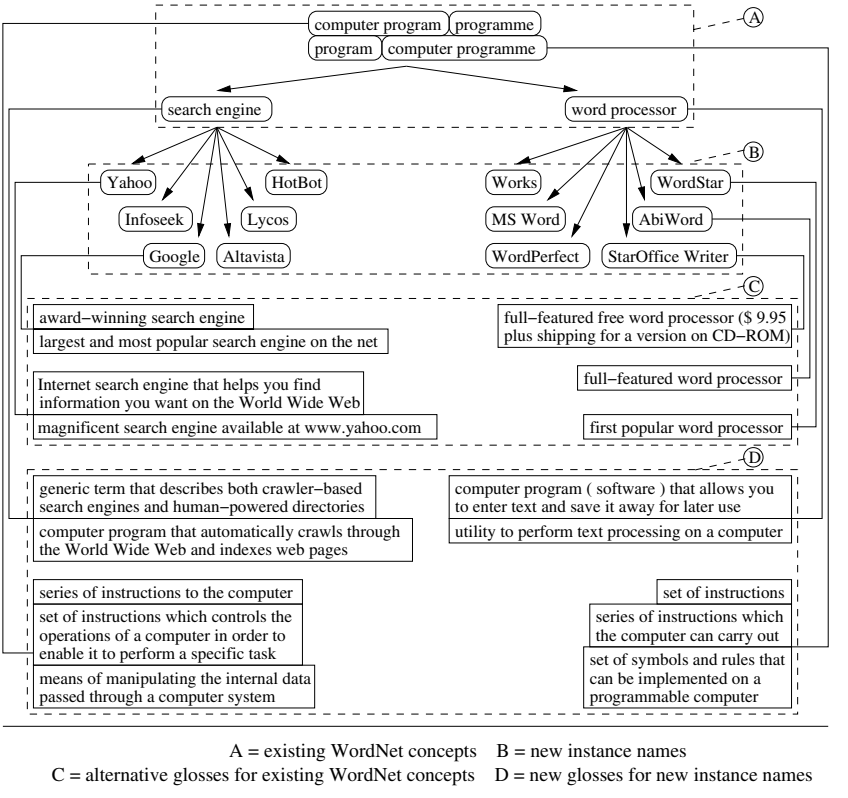

Fig. 1. Overview of augmenting WordNet with instance names and alternative glosses extracted from Web text nuggets

and robust heuristics. The overall method is data-driven, without any a-priori restrictions on the type of the targeted concepts and glosses. Thus, it harnesses some of the unstructured knowledge available across the Web.

3 The Web as a Source of Factual Text Nuggets

The goal of the Web is sharing information and knowledge. The use of complex text processing tools as a step towards accessing the knowledge within the text is impractical. Without attempting true text understanding, it is still possible to extract a small part of the available knowledge via shallow text processing. In this case, the knowledge is assumed to be encoded within text nuggets.

3.1 Text Nuggets

A text nugget captures a factual property of a lexical concept (phrase or word). Both the nugget and the lexical concept occur in text. The type of property

Table 1. Examples of Web sentences containing descriptive (*Desc*) and categorical (*Categ*) nuggets (W=lexical concept; X=text nugget)

Desc	[WordNet]^W is an [English lexical reference system based on current psycholinguistic theories of human lexical memory]^X.
Categ	[The world is less dominated by mega-cities (10 million or more people)]^X such as [Mexico City]^W, than many predicted only a few years ago [..]

encoded in the nugget determines its form and relation with the concept. The presence of potential concepts and simple domain-independent, lexico-syntactic patterns in sentences is a signal of a text nugget associated with the concept. A *descriptive* nugget introduces distinguishing properties (differentia) of the lexical concept W to which it is associated, by connecting it to a description X. As shown in Table 1, descriptive nuggets often occur in the form of appositives, linking verbs and subordinate clauses. The patterns are encoded as:

(1) ⟨W [,] [who|which] [is|was] [the|a|an] X [,|.]⟩
(2) ⟨[StartOfSent] W [,] [a|an|the|who|which] X [,]⟩
(3) ⟨[StartOfSent] [A|An|nil] W [is|was] [the|a|an] X [,|.]⟩

A *categorical* nugget, as shown in Table 1, is likely to provide the category (genus) of the associated concept. The set of patterns can be summarized as:
(4) ⟨[StartOfSent] X [such as|including] W [and|,|.]⟩.

3.2 Extraction of Text Nuggets

To ensure robustness on large collections, the extraction relies on lightweight tools and minimal resources. As a pre-requisite, the input documents are first pre-processed to filter out HTML tags. After tokenization and sentence-boundary detection, documents are part-of-speech tagged using the TnT tagger [7]. Each of the lexico-syntactic patterns is matched against document sentences, resulting in pairs (W,X) of a concept and an associated text nugget for each match. There are two modes of operation, depending on how the concepts W are detected:

1. The concepts are part of a *closed vocabulary* (e.g., WordNet nouns) which is given as input. In this case, their detection is equivalent to searching the current sentence for the longest matching vocabulary entries that are not preceded by noun modifiers (other nouns or adjectives), then checking for the presence of a pattern around them.
2. The concepts are part of an *open vocabulary*, which is not specified as part of the input. Since languages such as English tend to distinguish proper names from other nouns through capitalization, each sequence of capitalized terms in the sentence is marked as a potential concept. Non-capitalized sequences (complex noun phrases) are not considered due to increased difficulty in detecting their boundaries.

In both cases, potential concepts that are not associated with a text nugget via a pattern are discarded. The output is a set of concepts with their corresponding nuggets (descriptive or categorical) as derived from Web documents.

4 Derivation of Higher-Level Information

Text nuggets represent low-level information that is not directly suitable for existing resources. This section describes the processing of descriptive and categorical nuggets to derive alternative glosses and new instance names respectively.

4.1 Alternative Glosses

Given a lexical concept, its descriptive nuggets define a semantic space that is usually only partially overlapping with that defined by the corresponding WordNet gloss. The partial (vs. complete) semantic overlap is mainly due to two reasons. First, a nugget may reveal properties that are different (although useful) and therefore not directly comparable to those included in the manually-created WordNet gloss. Second, a nugget may include a "perfect" definition, but for a sense that is different from the one(s) in WordNet. Examples are *Jaguar*, which is found as an *animal* in WordNet but also as a *car company*, *operating system codename*, and *64-bit video game system* in the nuggets extracted from the Web; and *Metropolis*, which is a *city* or the *people living in a city* in WordNet, but also a *movie*, *algorithm*, *club*, *magazine* and *festival* in the extracted nuggets. Therefore, the main issue in converting the descriptive nuggets into alternative glosses is how to divide the semantic space defined by the set of nuggets.

The solution proposed here is to perform hierarchical agglomerative clustering [8] of the descriptive nuggets of a given lexical concept, based on pairwise nugget similarities. A practical metric for the similarity of two nuggets is the dot-product of their term vectors, after removal of stop words and vector Euclidean-length normalization. The initial weights are the term frequencies within the nugget. The first few (three, in this case) non-stop terms in the nugget are heuristically assigned higher weights (three times higher), following the intuition that they correspond frequently to the genus of the lexical concept. In terms of clustering method, our early experiments suggest that group-average clustering [8] is the best for our purpose. The method starts by placing each nugget into a separate cluster, and builds hierarchical clusters iteratively. All elements (nuggets) of intermediate clusters contribute to the computation of inter-cluster similarities, as a group-average of the pairwise element similarities. The clustering ends when all inter-cluster similarities are lower than a minimum threshold, which is experimentally set to 0.1. The gloss clusters are ranked in decreasing order according to their number of elements as illustrated in Table 2.

4.2 New Instance Names

The process of deriving instance names uses categorical nuggets, which are searched for the noun phrase that encodes the category of the associated lexi-

Table 2. Examples of top-ranked gloss clusters and some of their elements (R/S=rank/size of the cluster)

R/S	*Examples of glosses in the cluster*
Thomas Jefferson:	
1/21	third president of the United States of America
	third president of the US
	3rd president of the US and the author of the Declaration of Independence
2/17	author of the Declaration Of Independence was educated at William and Mary College
	key author of the Declaration of Independence
	principal author of the Declaration of Independence
3/10	best educated and the most original man of his day
	noted scientist and inventor himself
Joshua Tree:	
1/7	fascinating place to visit and photograph
	enchanting place to go for a hike
	great place to see the desert largely unspoiled by vehicles
2/5	big national park in the desert outside LA
	namesake of Joshua Tree National Park near Palm Springs
	home of Joshua Tree National Park and situated in the very heart of the Morongo Basin
3/2	is actually a type of yucca
	variety of yucca and a member of the Lilly family
search engine:	
1/36	computer program that searches the indexes of web sites using keywords
	computer program that automatically crawls through the World Wide Web and indexes web pages
	computer program that searches a database to find those objects that meet the search criteria you specify
2/29	web site that is linked to a database of web sites
	web site that is devoted to searching all the other web sites
	Web site that is like a catalog of the Web

Table 3. Samples of categories and their top instance names acquired from the Web

Category	*Top instance names*
color	Black, Red, White, Blue, Green, Yellow, Orange, Pink, Purple
rapper	Eminem, Jay-Z, Nas, Dmx, Snoop Dogg, Dr. Dre, Ja Rule
high-speed network	ATM, Gigabit Ethernet, B-ISDN, FDDI, Myrinet, Frame Relay
operating system	Linux, Windows, Windows NT, Unix, DOS, Solaris
car rental company	Hertz, Alamo, Budget, Avis, National, Dollar, Thrifty, Europcar

cal concept. This phrase is approximated by the rightmost non-recursive noun phrase whose last component is a plural-form noun, e.g. *mega-cities* for *Mexico City* in one of the entries of Table 1. Such a coarse approximation is more scal-

able to millions of Web documents. The acquisition of instance names with their categories from the Web is described in more detail in [9]. Table 3 illustrates instance names extracted in the open-vocabulary mode of operation. Note that WordNet may contain all new instances of a category (e.g., for *colors*), contain none of them (e.g., for *rappers*), contain only some of them (e.g., for *operating systems*), or not contain the category in any of its entries (e.g., *high-speed network* and *car rental company*).

4.3 Integration into Existing Resources

The insertion of bits of information extracted through shallow text processing from a decentralized, anonymized knowledge repository (the Web) into a high-quality hand-made resource (WordNet) is certainly challenging. Ideally, human intervention should not be needed for double-checking, correcting or guiding the integration process. But this comes at odds with the need to insure the correctness or at least graceful degradation of the resulting resource.

A conservative integration approach will embed new knowledge into WordNet while minimizing the chances of errors. In the case of alternative glosses, this translates into applying a set of restrictive filters to any gloss before linking it to an existing WordNet lexical concept. First, the gloss must belong to a relatively higher-ranked, that is, larger gloss cluster of that word as shown in Table 2. Second, the gloss must have a relatively high similarity with a WordNet gloss of that word. The metric for computing the similarity of a gloss to a reference WordNet gloss is the same as that used for clustering, namely the dot-product of the non-stop, length-normalized term vectors. Table 4 shows the most similar alternative glosses extracted for a subset of existing WordNet concepts.

New instance names correspond to a new node being linked to an existing node at the bottom of the hierarchies. For example, each of the instance names *Google* (in the category *search engine*), *Swiss National Bank* (in *central bank*) and *Joschka Fischer* (in *foreign minister*) generates a new leaf node inserted under the WordNet concepts *search engine*, *central bank* and *foreign minister* respectively. A different set of conservative restriction filters applies here. First, there should be only one possible insertion point, i.e. the category of the name matches exactly one WordNet concept, and the latter is a leaf node. If a category does not match any WordNet concept, its modifiers are discarded until a match is found. Thus, *high-level programming languages*, *Internet portals* and *science fiction writers* match the WordNet concepts *programing language*, *portal*, and *writer* respectively. It is useful to assign glosses to new instance names as well. In this case, an additional conservative restriction is that the gloss must contain the word to which it is linked. For example, if *Google* is inserted under *search engine*, its gloss must contain a lexicalization of *search engine*.

With the conservative approach discussed so far, the restriction filters remove possible errors due to spurious extraction or ambiguity, at the expense of discarding a lot of nuggets that might be otherwise useful. A more aggressive approach gradually removes restrictions. This increases the percentage of new knowledge that can be integrated into WordNet. For example, instance names with several

Table 4. Examples of reference WordNet glosses (*Ref*) with their most similar alternative glosses (*Alt*) extracted from the Web

Ref	(Apollo): Greek god of light; god of prophesy and poetry and music and healing; son of Zeus and Leto; twin brother of Artemis
Alt_1	Greek god of light and music
Alt_2	Greek god of death and pestilence as well as of the sun and medicine
Ref	(chemistry): the science of matter; the branch of the natural sciences dealing with the composition of substances and their properties and reactions
Alt_1	branch of science that deals with the composition and properties of matter
Alt_2	science dealing with the structure and composition of substances and the mechanisms by which changes in composition occur
Ref	(artificial intelligence): the branch of computer science that deal with writing computer programs that can solve problems creatively; "workers in AI hope to imitate or duplicate intelligence in computers and robots"
Alt_1	branch of computer science that involves writing computer programmes that solve problems creatively
Alt_2	branch of computer science concerned with making computers think
Ref	(fluoxetine): a selective-serotonin reuptake inhibitor commonly prescribed as an antidepressant (trade name Prozac)
Alt_1	Selective Serotonin Reuptake Inhibitor (SSRI)
Alt_2	selective serotonin reuptake inhibitor (SSRI) has been suggested in reducing irritability in PTSD patients
Ref	(emoticon): a representation of a facial expression (as a smile or frown) created by typing a sequence of characters in sending email; ":-(and :-) are emoticons"
Alt_1	face created out of keyboard characters
Alt_2	group of 3 or 4 punctuation characters which resemble a face turned sideways displaying a mood

possible points of insertion could be reviewed, rather than discarded. Similarly, a human reviewer might inspect gloss clusters that are not very similar to the reference WordNet gloss, yet describe a different, valid, relevant property of that concept; alternatively, such a gloss cluster could reveal a different word sense that is absent from WordNet. Human intervention could also refine the linking of an alternative gloss, by selecting the precise sense of the word to which the gloss applies. Moreover, one could explore the idea of creating missing intermediate concepts, rather than preserving the WordNet hierarchy structure. Intermediate concepts occur as categories of extracted instance names, e.g. *software company* and *high-level programming language*. They would fit under existing concepts (*company* and *programming language*), but are missing from WordNet.

5 Evaluation

5.1 Experimental Setting

The experiments are performed on approximately 300 million Web documents in English from a snapshot of the Google index from 2003. Two parallel runs

Table 5. Words with the highest number of descriptive nuggets

Word	*Count*	*Word*	*Count*	*Word*	*Count*	*Word*	*Count*
Trauma	26936	Jesus Christ	6791	Church	4055	John	2954
God	18561	Jesus	5148	New York	3883	Tigers	2934
Christ	11638	Internet	5110	Holy Spirit	3501	commission	2870
United States	9459	Spirit	4646	President	3406	President Bush	2851

use this data. Run_1 aims at extracting alternative glosses for existing WordNet concepts. After discarding words not starting in alphabetic characters, the closed vocabulary contains 114,487 WordNet noun entries. Run_2 collects new instance names and their glosses using an open vocabulary.

5.2 Results

The extraction method identifies one or more descriptive nuggets in the collection for 60% percent of the input words in Run_1. Among the 46,329 words without a descriptive nugget, 3,598 are compound nouns starting in *"genus"* (*genus icterus*, *genus iguana*), 1,136 start with *"family"* (*family adelgidae*, *family Erethizontidae*), and 324 start with *"order"* (*order anoplura*, *order batrachia*). Some of these words are synonyms to words which have extracted nuggets. Thus, *family Erethizontidae* is a synonym of *Erethizontidae*, whose extracted nugget *"New World porcupines"* is very similar to WordNet's *"New World arboreal porcupines"*.

The average number of descriptive nuggets across the words with at least one such nugget is 110, whereas the median is 11. Table 5 shows the words with the highest number of descriptive nuggets. We were baffled to see *Trauma* at such a high rank, so we checked a sample of its nuggets. All are due to adult-content spam, and have the form *"Trauma, which is a blow to ⟨SpamPhrase⟩, can occur under a variety of circumstances"*. In fact, the majority of the URLs from which they were extracted were available in 2003 but are no longer valid. Three words from Table 5 illustrate another undesirable phenomenon, namely nuggets that refer to particular instances of the same, more general concept (a certain *commission*, one *John*, a certain *president* etc.). This problem is related to the correct identification and disambiguation of names in text [10].

The quality of the descriptive nuggets in Run_1 is further investigated through manual evaluation of a set of 100 randomly selected WordNet nouns with exactly one descriptive nugget extracted from the Web. Nuggets are deemed as *partially* correct if they contain enough information to infer the genus but little else. Other nuggets, which are valid but apply to a different sense of the word than the senses in WordNet, are evaluated as a separate category, similarly to the evaluation in [11]. An example of a word with a different sense is *Leonberg*, whose WordNet definition captures the sense of a large dog breed rather than that of a city. Table 6 shows that most nuggets are correct or partially correct.

A complementary evaluation of Run_1 considers a different set of 100 randomly selected WordNet words, without any restriction on the number of their descriptive nuggets. For each word, the alternative glosses are compared against

Table 6. Accuracy on a random set of 100 words with only one alternative gloss extracted from the Web (C=correct; P=partially correct, i.e., correct but incomplete; O=other sense than WordNet; I=incorrect)

Type	*Pct.*	*Examples*	
		Word	*Alternative Gloss*
C	35%	Colubridae	advanced snakes and the largest snake family
		fire warden	voluntary officer responsible for safe rural fire management within the community
P	35%	Gentiana acaulis	stemless gentian
		Polygalaceae	family of plants
O	3%	Leonberg	small suburb outside Stuttgart
		Rumex	old Latin word for "lance"
I	27%	siege of Vicksburg	city's main claim to fame
		yobbo	backwards spelling of boy

the WordNet reference glosses of that word. The comparison is implemented through the dot-product similarity of the length-normalized term vectors, as discussed in Section 4.3. In the experiment, the alternative glosses whose similarity value is below 0.5 are discarded; the rest are manually checked. Out of the 100 test words, 41 have at least one extracted descriptive nugget above the threshold. The average number of such nuggets per tested word is 7, and the median of 4. The nuggets are classified into one of the judgment classes listed in Table 6. Among the 280 verified nuggets, 195 nuggets are deemed correct, 70 partially correct, 6 incorrect and 9 correct for a different word sense. Nuggets like *"branch of computer sciences"* (for *artificial intelligence*), *"German composer"* (for *Ludwig van Beethoven*) and *"acronym for: Light Amplification"* (for *laser*) are marked as partially correct rather than correct. Note that the decision on whether a nugget is only partially correct is highly subjective. Depending on the application, one could argue that the classification of *"German composer"* as partially correct is pessimistic, since the nugget contains both the genus (*composer*) and what may be relevant differentia (a composer from *Germany*).

To assess the impact of the instance names on WordNet, it is useful to look at the categories to which they are associated according to the data. There are almost 300,000 such raw, distinct lexicalized categories collected in Run_2 that are not isolated occurrences, i.e. are associated with at least 4 instance names. As a primary aggregated result, around 15,000 of these data-driven categories are WordNet nouns. Under ideal conditions (perfect extraction, non-ambiguity, sense matching the one present in WordNet etc.), all instance names associated to these categories potentially belong under existing WordNet concepts.

A secondary result of Run_2 is the acquisition of interesting categories which may be missing from WordNet (see Section 4.3). Intuitively, a category containing a larger number of instances is a better candidate to become a new concept in WordNet. A few of the larger extracted categories with the head *company* are already in WordNet, e.g. *insurance*, *pharmaceutical* and *oil (company)*; others, like

technology, *software*, *blue chip*, *media*, *high-tech*, *Internet*, *telecommunications*, *manufacturing*, *industrial* and *biotechnology (company)* are not in WordNet. Similarly, among the larger categories with the head *programming language*, the category *object-oriented (programming language)* is already in WordNet, whereas none of *web*, *high-level*, *logic*, *functional*, *Internet*, *procedural*, *imperative*, *general purpose* or *structured (programming language)* belong to its noun database.

6 Discussion and Previous Work

One of the early proposals for mining unstructured text with lightweight extraction patterns was considered precisely in the context of discovery of WordNet-style information, i.e. hyponyms [12]. Others successfully applied lightweight patterns to text collections for various applications, including summarization [13], information extraction [14] and question answering [15, 16]. As the availability of large text corpora increased, it became possible to collect large semantic lexicons and resources of instance names, either manually or automatically [17]. However, these experiments tend to organize the instance names into a fixed, coarse-grained set of categories. Comparatively, the new instance names collected in this paper are associated to a large set of data-driven categories. The usefulness of a larger and noisier resource, namely the Web, is indicated by experiments in finding domain-specific definitions [18] and encyclopedic term descriptions [11]. However, we are not aware of work that extends existing WordNet glosses with alternative glosses extracted from the Web, for existing concepts (closed vocabulary) and unspecified instance names (open vocabulary). Alternative glosses are also a possible source for paraphrases, whose acquisition is different from recent approaches focused specifically on collecting paraphrases [19].

7 Conclusion

The largest search engines provide access to more than 8 billion Web documents, as part of an unstructured, unreliable yet powerful knowledge resource that seems to be growing endlessly. Hidden inside documents on different topics, small text nuggets capture some information about the world in a form that is relatively easier to exploit automatically. This paper described a lightweight method to collect text nuggets from the Web and morph them into information that can be linked into existing lexical, hierarchical resources. The insertion of automatically derived glosses and instance names into a resource such as WordNet is certainly challenging. Yet leveraging formally-structured, human-validated resources, on one hand, and data-driven sets of instance names and definitions on the other, opens the path to new applications of the reloaded resources.

References

1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press (1998)

2. Agirre, E., Rigau, G.: Word sense disambiguation using conceptual density. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Copenhagen, Denmark (1996) 16–22
3. Chai, J., Biermann, A.: The use of word sense disambiguation in an information extraction system. In: Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), Menlo Park, California (1999) 850–855
4. Dorr, B., Katsova, M.: Lexical selection for cross-language applications: Combining LCS with WordNet. In: Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, Pennsylvania (1998) 438–447
5. Green, S.: Automatically generating hypertext in newspaper articles by computing semantic relatedness. In: Proceedings of the 2nd Conference on Computational Language Learning (CoNLL-98), Sydney, Australia (1998) 101–110
6. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico (2003) 805–810
7. Brants, T.: TnT - a statistical part of speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00), Seattle, Washington (2000) 224–231
8. Voorhees, E.: Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Information Processing and Management **22** (1986) 465–476
9. Paşca, M.: Acquisition of categorized named entities for Web search. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM-04), Washington, D.C. (2004)
10. Wacholder, N., Ravin, Y., Choi, M.: Disambiguation of proper names in text. In: Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97), Washington, D.C. (1997) 202–208
11. Fujii, A., Ishikawa, T.: Summarizing encyclopedic term descriptions on the web. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland (2004) 645–651
12. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), Nantes, France (1992) 539–545
13. Schiffman, B., Mani, I., Concepcion, C.: Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-01), Toulouse, France (2001) 450–457
14. Phillips, W., Riloff, E.: Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02), Philadelphia, Pennsylvania (2002) 125–132
15. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02), Philadelphia, Pennsylvania (2002)
16. Solorio, T., Pérez, M., Montes, M., Villasenor, L., López, A.: A language independent method for question classification. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland (2004)

17. Cucerzan, S., Yarowsky, D.: Language independent named entity recognition combining morphological and contextual evidence. In: Proceedings of the 1999 Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), College Park, Maryland (1999) 90–99
18. Liu, B., Chin, C., Ng, H.: Mining topic-specific concepts and definitions on the web. In: Proceedings of the 12th International World Wide Web Conference (WWW-03), Budapest, Hungary (2003) 251–260
19. Dolan, W., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland (2004)