Knowledge extraction from machine readable
dictionaries (MRDs) and introduction to existing lexical resources

Hugo Gonçalo Oliveira
CISUC, University of Coimbra, Portugal

1 Introduction

This document starts by telling the ”story” of exploration and utilisation of machine readable dictionaries (MRDs) as a source of knowledge and then introduces some known lexical resources.

2 Knowledge extraction from MRDs

This section is about using machine readable dictionaries (MRDs) as a source of knowledge. It started more than 30 years ago but there is still research going on that explores them in order to extract and organise knowledge.

2.1 An introduction to MRDs

Machine Readable Dictionaries (MRDs) are electronic versions of dictionaries, especially designed to be used by or through machines. MRDs are usually stored in a database and can be queried via some interface.

The first dictionaries known to have a machine-readable format were Merry Webster’s Pocket Dictionary (MPD) and Webster’s Seventh New Collegiate Dictionary (W7)¹ that were manually keyboarded and distributed in multiple reels of magnetic tape, back in the 1960s [1]. From that time, the creation of electronic versions of the dictionaries had in mind helping natural language processing (NLP) systems.

Besides the aforementioned MRDs, the electronic version of the Longman Dictionary of Contemporary English (LDOCE)² also started to be explored, later in the 1980s, with the purpose of evaluating how useful it could be for NLP [2]. An LDOCE entry may contain two kinds of codes that revealed to be very helpful concerning semantic information extraction and word sense disambiguation (WSD): box codes and subject codes, both organised into hierarchies. While the former are a set of primitives to assign type restrictions on nouns,

¹ The current version of the W7 is available for online search in http://www.merriam-webster.com/
² The current version of the LDOCE is available for online search in http://www.ldoceonline.com/
adjectives and on the argument of verbs (primitives like abstract, animate or human, conforming the classical notion of the IS-A relation), the subject codes consist of headings and sub-headings that classify the words by subject (terms like engineering or economics) [3].

2.2 In the beginning

Back in the 1970s and through the 1980s, MRDs started to be the target of empirical studies in order to assess the possibilities of using them as a source of semantic knowledge, useful for NLP [4]. The work of Nicoletta Calzolari includes the exploration of the definitions in order to organise the dictionary into a lexical database (LDB), where morphological and semantic information about the defined words could be obtained directly [5]. If the created database is well structured, it is easier to automatically identify some syntactic and semantic relations between the entries of the MRD.

Similar work took place for English when the electronic versions of the LDOCE and the MPD were used as a source of information to build such a structure. In 1980, Michiels et al. [2] explored the files of the LDOCE, presenting its structure and taking some conclusions about the properties of the definitions. Like other authors, they concluded that the vocabulary in a dictionary is very limited, easing its processing in order to obtain relations between syntactic or semantic structures.

In the same year Robert Amsler [6] explored the structure of the electronic version of the MPD. He noticed that the text of the definitions often consists of a genus and a differentia:

- The genus identifies the superordinate concept of the defined word. In other words, the defined word is a "type of" the genus and there is a (usually) hyponymy relation between the former and the latter.
- The differentia consists of the specific properties responsible for the distinction between the respective instance of the superordinate concept and other instances of the same concept.

If the genus is extracted and disambiguated, it is possible to build semantic hierarchies based on the hypernymy relation (for nouns) or troponymy (for verbs). The terms genus and differentia are used in most of the publications in this research topic.

Having in mind that it was possible to extract a huge amount of semantic information from the dictionary, Amsler proposed a taxonomy consisting of hierarchies of nouns and hierarchies of verbs. He called them tangled hierarchies and created them based on the analysis of the definitions in the MPD and on the manually disambiguated head of each definition [7]. The hierarchies were organised in a way that the most specific words could be found in the lower levels and the most generic (like cause, thing, class, being,...) in the top. Another conclusion taken by Amsler was that the dictionary contains at least two clear taxonomic relations: is-a (hyponymy) and is-part (part-of).
Taking advantage of the restricted and specific vocabulary and of the regular syntactical occurrences in a MRD, Calzolari [8] also suggests sets of patterns that are regularly used and examines the occurrence of the hyponymy and "restriction" relations. She claims that hyponymy is the most important and evident relation in the lexicon and can be easily extracted from an MRD with the identification of the genus of the definition.

Some years later, Markowitz et al. [9] identified a set of textual patterns that occur in the beginning of the definitions of the W7. The presented patterns imply relations between nouns, namely the superordination and the member-set relations; imply that the defined noun is a human being; and identify verbs or adjectives as active or stative. The following are some examples of the identified patterns:

- Superordination: any, any of;
- Member-set: member of;
- Human noun: one;
- Information about verbs in the definition of nouns: act of <active verb>ing, the act of <stative verb>ing, the state of being <adj>;
- Adjectives: of or relating to (stative), being (active).

2.3 First (semi) automatic procedures for processing MRDs

In 1985, Chodorow et al. [11] proposed two "head-finding" heuristics to identify the genus of a definition: one for nouns and another for verbs. Bearing in mind the structure of the definitions and assuming that a defined concept is often an hyponym of its superordinate concept, he took advantage of the restricted vocabulary used in the definitions to developed semi-automatic recursive procedures aiming the extraction and organisation of semantic information into taxonomic trees. The definitions did not have to be completely parsed due to their predictability. However, the human user played an important role when it came to WSD. The authors claim a virtual 100% accuracy in the genus extraction for verbs, using a very simple heuristic: the head is the single verb following the word to. If there is a conjunction of verbs following to, they are all heads. For example:

- winter (v): to pass the winter → pass
- winter (v): to keep, feed or manage during the winter → keep, feed, manage

When it comes to nouns, the task is much more complex due to their greater variety, but they could still take advantage of the special and predictable style of their definitions and came up with an heuristic for the extraction of the genus. The heuristic is based on the isolation of the substring containing the head, which is bounded on the left by a word like a, an, the, its, two, three, ..., twelve, first, second, ... and is bounded on the right by a word like:

- a relative pronoun (introducing a relative clause);
– a preposition not followed by a conjunction (thus, introducing a complement to the head noun);
– a preposition-conjunction-preposition configuration (also introducing a complement);
– a present participle following a noun (thus, introducing a reduced relative clause).

After isolating the substring containing the head, the search for the head begins. It is typically the rightmost noun in the substring. Chodorow et al. claim 98% accuracy for the heuristic for nouns.

Hiyan Alshawi analysed the definitions of the LDOCE where syntactic patterns were identified to make possible the construction of semantic structures based on the meaning of the defined words [14]. These structures were derived from the identification of the subordinated terms or modifiers, prepositions and other words that could indicate relations in the definition. A set of semantic relations (e.g. class, purpose, manner, has-part) and, in some cases, specific properties were extracted and included in the semantic structures. Alshawi [15] also proposed a specific grammar for the derivation of the definitions of the LDOCE. His main concern was to accomplish partial syntactical derivation based on the structure of the definitions of this specific dictionary, so the application of the grammars to unrestricted text or to other dictionaries might not be a good option.

2.4 Typical problems

One of the first noticed problems when using dictionaries to build a taxonomy is circularity, often present in dictionary definitions [10]. This phenomenon occurs when starting in an entry, we go to the entry corresponding to the head of its definition and, eventually after several levels of recursion, we end up in some entry that we had already been at. The following definitions constitute a made up example of circularity:

– portion - a part of a whole;
– part - a piece of something;
– piece - a portion of some material;

In Amsler’s [7] work circularity is referred to as loops (groups of words defined in a circular way) and the importance they have is discussed. He believes that loops are usually the evidence of a truly primitive concept, like for example the set containing the words CLASS, GROUP, TYPE, KIND, SET, DIVISION, CATEGORY, SPECIES, INDIVIDUAL, GROUPING, PART and SECTION.

When identifying the genus term to obtain hyponymy relations, attention should be paid to certain head words that give special information about the defined word and can be related with other types of relation. Chodorow et al. [11] called them “empty heads” and gave specific examples (one, any, kind, class,
manner, family, race, group, member, ...). To deal with this problem, whenever his procedures met an empty head, the noun word following the preposition of (as in kind of boat) was interpreted as the head. Although it seemed reasonable, Guthrie et al. [12] argue that since some of the words Chodorow et al. considered to be "empty" are usually related with other relations like, for instance, the word member which is related with a member-set relation [9] or the word part which is related with a is-part relation (included by Amsler [7] in his tangled hierarchies). Nakamura and Nagao [13] provide a list of function nouns that appear in dictionary definitions and the relations they are usually associated with:

- kind, type → is-a
- part, side, top → part-of
- set, member, group, class, family → membership
- act, way, action → action
- state, condition → state
- amount, sum, measure → amount
- degree, quality → degree
- form, shape → form

Another typical issue is the association of words that appear in the definition with their sense in the dictionary. For instance, the disambiguation of the genus is needed for the extraction of a taxonomy from a MRD. One of the biggest limitations of Amsler’s [7] and Chodorow et al.’s [11] work, is that the disambiguation of the genus requires human intervention. Bruce and Guthrie [3] worked on an automatic procedure to accomplish this task, that involves two subproblems:

1. identification of the genus/hypernym word from a definition;
2. disambiguation of that word into a concept;

While effective methods had been presented to solve the first problem the second one is more difficult. However, an algorithm was developed for the disambiguation of the genus, taking advantage of the box codes, subject codes and frequency of utilisation associated with each entry of the LDOCE. The algorithm, which the authors claim to have 80% accuracy, is stated as follows:

1. Choose the genus sense with the same semantic category as the headword, or with the closest more general category.
2. If there is a tie, choose the sense with the same pragmatic code.
3. If a tie remains or no genus sense meets the above criteria exists, choose the most frequently used sense of the genus word.

2.5 Broad-coverage parsing in MRDs processing

After some discussion about the advantages and the drawbacks of using string patterns or structural patterns to extract semantic information contained in the
definitions, Montemagni and Vanderwende [17] concluded that although string patterns are very accurate for identifying the genus, they cannot capture the variations in the differentia as well as structural patterns, and they proposed the use of a broad-coverage grammar to parse the dictionary definitions in order to obtain rich semantic information. String patterns are based on specific textual constructions of the definitions and were used by Chodorow et al. [11], Markowitz et al. [9] and others, while structural patterns are based on the analysis of the structure of the syntactic trees of the definition. Previous work on the automatic extraction of relations from MRDs (using string patterns) reported very good results, but this work focused mainly in the hypernym extraction. When it comes to the extraction relations depending on the differentia, string patterns have several reported limitations:

- When there is an enumeration of concepts at the same level: to make laws, rules or decisions;
- When there are parentheses in the middle of the definition;
- When it is necessary to identify functional arguments;
- When there are specific relations inside the definition: in pianta erbacea com bacche di color arancio the color feature should not be extracted as a feature of the defined word.

In spite of seeming an overkill to use a broad-coverage parser for definition text, the authors make the point that there are cases (relative clauses, parenthetical expressions, and coordination) when its use is warranted. The following is an example of an heuristic for the extraction of the purpose relation: if the PP with for is not a post-modifier of a verb used, then a purpose relation between the defined word and the head(s) of the PP can be hypothesised if the nearest noun that the PP post-modifies is the genus term.

Although dictionaries have been explored for several purposes, such as parsing, deriving semantic structures or WSD, to our knowledge they have not been converted into an independent resource of its own before the late 1990s (after several publications in that direction [20, 21, 22, 23]), when MindNet [24] was presented, which therefore can be said to be a sort of independent lexical ontology in a way that previous work was not.

William Dolan et al. [20] describe a strategy to build a structure of lexical knowledge automatically from LDOCE. Their approach uses a broad-coverage parser to process the dictionary entries. Using a broad-coverage parser avoids the need to adapt it for different MRDs. The authors state that much information about a word can be found in the definition of other words. Looking for some word in the definitions of other words makes it possible to obtain many relations that include the first word. In order to create a semantic network, the set of relations to extract must be defined. The definitions are then parsed and searched for patterns that imply semantic relations. Each identified relation is added to the sense entry in a semantic structure representing the definition. The

---

3 prepositional phrase
resulting network contains words linked by means of semantic relations. Inferencing over the obtained network can be done to resolve semantic ambiguities on text.

William Dolan [21] also worked on a heuristic approach to automatically identify related senses of the same word in a dictionary, where each word can have definitions divided into more than one sense. Dolan’s approach consisted of identifying which senses are semantically related and which ones are fundamentally different, offering benefits for semantic processing and for the mapping of word senses across multiple MRDs. This process was called word sense “ambiguation”.

Lucy Vanderwende [22] presents an algorithm for the automatic interpretation of noun sequences in unrestricted text. Her system uses broad-coverage semantic information, acquired automatically by analysing the definitions in an on-line dictionary. Vanderwende [23] also worked on treatment of lexical ambiguity present in the language used by on-line dictionaries. A dictionary is processed multiple times, each time refining the lexical information previously acquired and identifying new information.

In his PhD thesis, Stephen Richardson [25] discusses the creation of the lexical knowledge base (LKB) that would be known as MindNet. In order to achieve his goal, the entries of the LDOCE are converted into a more formal representation, resulting in a dictionary called MIND (Microsoft Natural Language Dictionary). The syntactical trees of the definitions are obtained with the help of a broad-coverage parser and are then transformed into a logical form. After this, a set of heuristics is applied in order to convert the logical form into relational form, where semantic relations are clear. The relations that include a word can be obtained from all the definitions where that word occurs and the resulting structure can be inverted, giving rise to a lot more relations. With the resulting resource it is possible to browse for relation paths (see Section 3.5). Similarity of words can be inferred and other conclusions can be taken from the analysis of the paths.

In 2005, O’Hara [28] wrote about the empirical extraction of semantic relations from dictionaries. Special attention was given to the information in the differentia to find distinctions between co-hyponyms to accomplish WSD. Relations like used-for or has-size can be used to learn important information about the concepts. Dictionaries follow lexicography rules that ease the extraction of this kind of knowledge, however definitions are always incomplete or vague when it comes to certain details needed to understand some concept. In his studies, O’Hara uses WordNet [30] as a simple dictionary. Although the usual would be to adapt the parser, all the definitions are pre-processed and transformed in order to be easily interpreted by a general parser. This is done because dictionary definitions are often given by sentence fragments that omit the defined word. For example, the definition for lock is ”a fastener fitted to a door or drawer to keep it firmly closed”. This entry is transformed into ”a lock is a fastener...”. There are different transformations, depending on the grammatical category of the words. In his approach, O’Hara used a broad-coverage dependency parser to determine
the syntactic relations present in a sentence. Then, the surface-level syntactic relations determined by the parser are disambiguated into semantic relations between the underlying concepts. Isolating the disambiguation from the extraction allows great flexibility over earlier approaches. After a disambiguation process the relations are weighted according to their relevance to the assigned concepts resulting in a labeled direct graph where each link has a probability attached. The network is then converted into a Bayesian network. The author believes that the Bayesian network representation of the differentiating information can be used to improve WSD systems that use both statistical classification as well as probabilistic spreading activation.

2.6 Critical work

MRDs are certainly an important source of knowledge about language and the world but their organisation does not favour their direct use as NLP tools, since they were created in order to be read by humans. Wilks et al. [16] mention three assumptions that should be made to accomplish the objective of automatically extracting knowledge from a dictionary, and transform MRDs into machine tractable dictionaries (MTDs):

1. **Sufficiency**: Determines whether the knowledge is strong enough and contains enough linguistic knowledge on the World to be the target of computational text processing.

2. **Extricability**: Determines whether it is possible to specify a set of computational procedures capable of extracting large scale semantic information from a MRD without human intervention, in a general format suitable for subsequent text analysis processes.

3. **Bootstrapping**: Addresses the initial linguistic knowledge needed to automatically extract knowledge from the definition texts.

The authors say that projects based on the manual construction of semantic structures have pessimistic visions concerning **Extricability** and **Bootstrapping**. Three methods based on the former assumptions are presented for the automatic extraction of knowledge. The methods differ in the amount of initial information needed:

- The first approach is based on co-ocurrences that permit the establishment of associations between words, without needing initial linguistic information.
- The second uses a grammar and a collection of linguistic patterns enabling, besides other items, the identification of the *genus* (hypernym) and the *differentia* for each entry in the dictionary.
- The last approach is the one that needs more initial knowledge, but permits the creation of a semantic structure free from circular references. Since circularity makes the knowledge too vague it is better to remove it. Starting with a set of 3,600 semantic units, corresponding to the various senses of the 1,200 words used to define the controlled vocabulary of the LDOCE, the algorithm analyses the additional words in the dictionary. Of the latter,
those whose definition uses words with an existing semantic unit lead to the generation of a new semantic unit of the entry. According to the authors, after four iterations all the words are processed.

Ide and Véronis [18] produced critical work about research on information extraction from dictionaries. The authors affirm that all the research done so far had not achieved significantly more than the extraction of small and limited taxonomies. Two problems concerning the information in dictionaries, that seems to be inconsistent and incomplete, are discussed:

- Dictionaries use inconsistent conventions to represent the knowledge in dictionaries. Work around the identification of the conventions turns out to be very time-consuming.
- The definitions are not as consistent as they should be. There are many variations to say the same thing. Dictionaries are the result of several lexicographers work for several years. Reviews and updates increase the probability of inconsistencies.

In different dictionaries (or sometimes, even in the same) there are definitions made up from hierarchies with very high levels and it is sometimes difficult to identify terms that belong to the same level.

The same authors [19] performed a quantitative evaluation of automatically extracted hypernymy relations. Hypernymy was chosen because it is the least arguable semantic relation and the easiest to extract. The authors believe that if the results for hypernymy are poor, they will be poorer for more complex domains and less clearly cut relations. The evaluation methodology consisted in comparing an "ideal" hierarchy, manually constructed, with hierarchies extracted from five dictionaries. The automatic extraction of hierarchies was based on the heuristics by Chodorow et al. [11] giving rise to tangled hierarchies that were later manually disambiguated. After inspection, it was noticed that these hierarchies had several serious problems:

- Incomplete information: some terms are (relatively randomly) attached too high in the hierarchy; some heads of definitions are not the hypernym of the defined word, but the "whole" that contains it; overlaps that should occur between concepts are sometimes missing.
- Difficulties at higher levels: all the heads separated by the conjunction or are considered to be hypernyms, but sometimes, when looking at the hierarchy, problems exist; circularity tends to occur in the highest levels of the hierarchy, possibly when lexicographers lack terms to designate certain concepts.

The authors state that hierarchies with these kind of problems are likely to be unusable in NLP systems and discuss means to refine them automatically. Merging the hierarchies of the five dictionaries and introducing "covert categories" drastically reduces the amount of problems from 55-70% to 6%. "Covert categories" are categories introduced specifically to include concepts that do not correspond to any particular word. For example, there is no word to describe the
hypernym of the concepts described by tool, utensil, implement and instrument, so a new "covert" hypernym is created.

2.7 Other approaches

Caroline Barrière's PhD thesis [26] presents a method for transforming a MRD for children into a LKB, made from Conceptual Graphs [27]. The American Heritage First Dictionary (AHFD) was chosen due to:

- Its limited size;
- The day-to-day knowledge and simple world knowledge included;
- The complete and simple sentence structure;
- Being a closed world because almost all the words used in the definitions are themselves defined;
- Bootstrapping capabilities, possible because of the closed-world system;
- Naive view of things (in contrast to an adult’s dictionary);
- Limited polysemy (limited number of senses for each word).

Conceptual graphs were used because they present a logic-based formalism and are flexible to express the background knowledge necessary for understanding natural language. Most of the structures used during the development of the LKB were based on this formalism. The usage of conceptual graphs allows the coexistence of ambiguous and non-ambiguous information in the LKB. All the definition sentences in the AHFD were transformed into conceptual graphs after being tagged, parsed, parsed to conceptual graph transformations, structurally disambiguated and finally semantically disambiguated giving rise to an automatically created type hierarchy. The LKB is then constructed using those graphs, exploring cluster formations and the expansion of the hierarchy using "covert categories". Concept clusters are large structures used to represent the meaning of a word by its interaction with other words. They consist of groups of words that help define each other.

It was concluded that "covert" classes should be included in the concept hierarchy. These unlabeled categories are often superclasses whose subclasses occupy the case relation to a verb (for example to live somewhere). This lead to different relations than the usual synonymy, hypernymy and meronymy.

In 2005, Eric Nichols et al. [29] introduced a system that automatically constructs ontologies by extracting knowledge from dictionary definition sentences. Their approach combines deep and shallow parsing of the definition sentences and generates semantic representation by the robust minimal recursion semantics (RMRS). For each definition, ontological relations are extracted from the most informative semantic representation. Using the deepest possible result, 81,582 relations were extracted from the Lexeed Semantic Database of Japanese and two evaluations were performed:

- An automatic evaluation consisting of the verification of the extracted relations in WordNet [30] and GoiTaikei (a manually created Japanese ontology).
The results for the relations obtained with the deep parsing had the best confirmation rate, 55.74%, and 63.31% if only nouns were considered. When it comes to relations obtained with the deepest result, the confirmation rates were 50.79% overall and 57.68% for nouns.

- A manual evaluation, consisting of a hand-verification of a set of the acquired relations. 88.99% accuracy is claimed.

WordNet and GoiTaikei seem to lack complete cover, since over half the relations were confirmed with only one resource. This might be what caused the difference between automatic and manual evaluation. The authors claim that their approach is easy to maintain and expand, because it requires few rules, and can be easily extended to cover any language with RMRS resources.

2.8 Summary
Table 1 puts side-by-side the attempts to extract and structure knowledge from MRDs referred in this section.

3 Lexical resources
In this section important existing lexical resources are presented. NLP capabilities of a language rely heavily on the existence of these resources, that can be seen as lexical ontologies. The following are examples of utilisation for them:

- Finding out the meaning of texts
- Inferring similarity of concepts
- Question & answering
- Machine translation
- Creative text production
- Studying the characteristics of a language

3.1 Princeton WordNet
Princeton WordNet [30] is a resource that combines traditional lexicographic information with modern computation in a dictionary based on psycholinguistic principles. It is freely available, widely used in computational linguistics and NLP and probably the most important reference when it comes to lexical ontologies in English. It was however manually created.

In the WordNet’s lexicon, the words are clearly divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the *synset*, which is a set of synonym words that can be used to represent one concept. The *synsets* are organised in a network of semantic relations, such as Hyponymy and Meronymy (between nouns) and Troponymy and Entailment (between verbs).

The original format used for the data representation of the Princeton WordNet seems to have several limitations [31] but different formats are available. There is for example a RDF/OWL representation of WordNet [32], published by the W3C consortium, that can be useful for Semantic Web applications.
<table>
<thead>
<tr>
<th>Work</th>
<th>MRD(s)</th>
<th>Relations</th>
<th>Extraction</th>
<th>WSD</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calzolari [10, 5, 8]</td>
<td>Italian Machine Dictionary (DMI)</td>
<td>Hyponymy, &quot;restriction&quot; or &quot;modification&quot;</td>
<td>Textual patterns matching</td>
<td></td>
<td>Tangled hierarchies</td>
</tr>
<tr>
<td>Amsler [6, 7]</td>
<td>MPD</td>
<td>Hypernymy, troponymy, part-of</td>
<td>Textual patterns matching</td>
<td>Manual</td>
<td>Tangled hierarchies</td>
</tr>
<tr>
<td>Markowitz et al. [9]</td>
<td>W7</td>
<td>Superordination, member-set, human, active/stative verb or adjective</td>
<td>Textual patterns (in the beginning of the definitions) matching</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alshawi [14, 15]</td>
<td>LDOCE</td>
<td>Class, purpose, manner, has-part</td>
<td>Syntactic patterns matching, a specific grammar for the LDOCE definitions</td>
<td>Relies on a LDOCE rule that only the most central senses of words appear in definitions</td>
<td>“Semantic structures”</td>
</tr>
<tr>
<td>MindNet [25, 24]</td>
<td>LDOCE</td>
<td>Hypernymy, Causation, Meronymy, Manner, Location and many more</td>
<td>Broad-coverage parser</td>
<td></td>
<td>Lexical Knowledge Base (MindNet)</td>
</tr>
<tr>
<td>Barrièrre [26]</td>
<td>AHFD</td>
<td>Oh, with, hypernymy, part-of, material, instrument, time, goal and more.</td>
<td>Conceptual graphs</td>
<td>Probabilities and heuristics</td>
<td>Lexical Knowledge Base</td>
</tr>
<tr>
<td>O’Hara [28]</td>
<td>Wordnet</td>
<td>Relations in the <em>differentia</em> (used-for, has-size)</td>
<td>Broad-coverage parser over pre-processed and simplified definitions</td>
<td>Yes...</td>
<td></td>
</tr>
<tr>
<td>Nichols et al. [29]</td>
<td>Lexeed Semantic Database of Japanese</td>
<td>Deep and shallow parsing combined</td>
<td>Hypernymy, synonymy, abbreviation, domain, other</td>
<td></td>
<td>Ontology</td>
</tr>
</tbody>
</table>

Table 1. Summary of attempts for knowledge extraction from MRDs.
3.2 EuroWordNet and MultiWordNet

EuroWordNet [33] and MultiWordNet [34] are two multilingual wordnets, created by two distinct models. The earlier consisted in trying to find correspondences between existing wordnets for different languages, while the later consisted of building language specific wordnets keeping as much as possible the semantic relations available in Princeton WordNet.

3.3 Cyc

Cyc [35] was not created in order to become a linguistic resource, but it is frequently cited and used in the NLP community. Opposing to WordNet, Cyc is highly formalised and all the knowledge is described with a language based in first order predicate logic.

Cyc is probably one of the richest knowledge resources and was one of the first projects on knowledge representation. Its authors claim Cyc is the world’s largest and most complete general knowledge base and commonsense reasoning engine.

OpenCyc is the open source version of Cyc. Its ontology is provided in OWL.

3.4 Portuguese wordnets

WordNet.BR [36] is a Brazilian version of the “WordNet concept” and started in 2002. Its database is structured around Synonymy and Antonymy manually extracted from a reference corpus where several dictionaries are included, and plans for adding more relations in the future have been reported in [36].

WordNet.PT [37] is another attempt of creating a Portuguese lexical resource from scratch, which started in 1999. The authors of WordNet.PT explicitly claim that the available resources for Portuguese NLP are not suitable for the automatic construction of such a resource. They use a set of 35 relations and are explicitly interested in cross-categorical relations such as those linking adjectives to nouns.

Despite having started some years ago, both of these projects seem to be still at earlier stages, possibly because the are both manually created and thus dependent on human effort.

3.5 MindNet

MindNet [24, 38] is a lexical knowledge base created by the Microsoft NLP research group. The resource was created by automatic tools, such as the broad-coverage parser MEG, used in the grammatical verification of Microsoft Word. This parser generates syntactical trees and logical where rules for the extraction of relations between words are applied.

MindNet is not a static resource. It represents a methodology consisting of a set of tools to acquire, structure, access and explore semantic information
contained in texts, so the semantic network was extracted not only from MRDs but also from encyclopedias, and other kinds of text.

MindNet contains a long set of relations, including Hypernymy, Causation, Meronymy, Manner, Location and many more. One interesting functionality offered by MindNet is the identification of “relation paths” between words. For example, if one looks for paths between car and wheel a long list of relations will be returned. The returned paths include not only simple relations like car is a modifier of wheel but also more complex ones like car is a hypernym of vehicle and wheel is a part of vehicle.

3.6 FrameNet

FrameNet [39] is another kind of lexical resource, which constitutes a network of relations between semantic frames, extracted from corpora and from a systematic analysis of semantic patterns in corpora. Each frame corresponds to a concept and describes an object, a state or an event by means of syntactic and semantic relations of the lexical item that represents that concept.

A frame can be conceived as the description of a situation with properties, participants and/or conceptual roles. A typical example of a semantic frame is transportation, within the domain motion, which provides the elements mover(s), means of transportation and paths and can be described in one sentence as: mover(s) move along path by means.

References


This article was processed using the \LaTeX macro package with LLNCS style