DIGITAL CULTURE
ITMO UNIVERSITY

Lexical and Semantic Resources in Automatic Text Processing

High School of Digital Culture
ITMO University
dc@itmo.ru

Contents

1 Introduction

Nowadays automatic text processing applications are enriched by resources that describe the knowledge about a language and the world using semantic networks of interrelated concepts. These resources should be large enough, so they can be used to great effect.

WordNet, a well-known English thesaurus, is an example of such resources. WordNet describes lexical relations between more than 150 thousand words and expressions. Lexical relations are used in applications to:

- Detect paraphrases (sentences conveying the same meaning).
- Identify the relationships between sentences in the text.
- Expand a query in information retrieval.
- Create question answering systems, etc.

We can use such lexical and semantic resources in hybrid approaches to multiple tasks by combining them with statistical and machine-learning methods, as the former contain ready-to-use knowledge.

Terminological resources coming in the form of semantic networks (thesauri, ontologies, or knowledge graphs) are also in demand in various domains that need to represent the domain knowledge. Biomedicine, for example, is especially interested in such resources.

This module will review the types of lexical relations that are used in automatic text processing, the problems related to representations of lexical meanings, as well as some well-known lexical and semantic resources.

2 Lexical Relations

Let's discuss the types of lexical relations that are in great demand in automatic text processing.

Synonyms are words that have the same core meaning but still differ in terms of their senses of the meaning since absolute synonyms are rare. Words that are synonymous with each other may share a denotatum. It means that the set of objects these synonyms refer to is the same for both of them. Words may have different connotations, for example (positive or negative associations that the words carry). The language registers of words may also vary (formal, informal, neutral, academic, etc.).

The following differences between synonyms are observed:

- Stylistic (wife/spouse)

- Expressive (horse/fleabag)
- Professional (lung inflammation/pneumonia)

Synonyms generally belong to the same part of speech. It's due to the standard test for synonymy performed by substitution of one word for another in a sentence. Syntactic synonyms (also known as derivatives) differ only in the lexical and grammatical form they have (for example, the verb verify and the noun verification).

Antonyms are words that have contrasting or opposite but not contradictory meanings. Antonyms fall within the following main categories:

- To start or to stop (begin/end)
- To act or to destroy the result (attach/detach)
- F or not F (wet/dry)
- More or less (wide/narrow)

One of the important types of relations is the relation between hyponym and hypernym, where hypernym is a word that has a broader meaning. The following tests are used to identify these relations. X is a hyponym of Y if the following two conditions are satisfied:

1. The statement 'A is Y' follows from the statement 'A is X'.
2. The statement 'A is X' doesn't follow from the statement 'A is Y'.

Here's an example: *It's a dog, therefore, it's an animal. It's an animal, but it doesn't mean it is a dog. It's a stallion, therefore, it's a horse. It's a horse, but it doesn't mean it is a stallion. It's scarlet, therefore, it's red. It's red, but it may not be scarlet.*

A part-whole relation (meronymy) is a set of somehow different relations rather than a clearly distinguished relation. We can formulate the definition of meronymy that however excludes some obvious cases of part-whole relations: *X is a meronym of Y if and only if the sentences of the form Y has Xs and an X is a part of Y sound acceptable for X and Y, when the noun phrases X and Y are considered generic concepts.*

The most central type of this relation is physical objects. Things that last for some time may have different forms changing over time. We refer to them as steps, phases, or stages. Such entities as groups, classes, and collections are in a meronymy relation with their elements, for example, tribe, team, committee, family, orchestra, court, squad, etc.

Lexical relations allow us to make conclusions based on texts. For example:

- No pets allowed => No dogs allowed (hyponym)
- Restaurant in Japan => Restaurant in Asia (Japan is a part of Asia; hence whole or holonym)
- Restaurant in Japan ⌒ Restaurant in China (co-hyponyms, i.e., hyponyms of the same hypernym)
- Good restaurant ⌒ bad restaurant (antonyms)

When answering questions about the text, we can use part-whole relations and co-hyponym relations as in the following example;

- When did Donald Trump visit France?
- The correct answer: Trump visited Paris in September (a geographic part).
- The information that is not considered the correct answer: Trump visited Spain in October (co-hyponym).

Word embeddings that are computed for a large amount of text data are widely used nowadays. Word embeddings make it possible to calculate the semantic similarity between words that correlates with human perceptions of semantic similarity. However, word embeddings fail to distinguish different types of relations between words with acceptable certainty. For example, word embeddings produce very high similarity between antonyms. However, sentiment analysis systems and question answering systems need to be sensitive to antonyms so they can produce correct results. Moreover, word embeddings can sometimes produce a high semantic similarity between the words that have different meanings, which is often difficult to explain.

3 WordNet-Like Thesauri

WordNet developed at Princeton University is a linguistic resource that links English concepts into a semantic network of meanings of words and lexical relations between them. WordNet is a free publicly available tool. It was used to conduct thousands of experiments in information retrieval and automatic text processing. WordNet 3.0 regroups more than 155 thousand lexemes and expressions into 117 thousand sets of synonyms (called synsets). The total number of lexeme-meaning pairs amounts to 200 thousand. Other countries are also developing resources for their languages based on the WordNet model.

The major WordNet relation is synonymy. Sets of synonyms (synsets) are basic structural elements of WordNet. The concept of synonymy is based on the criterion that two expressions are synonymous if the replacement of one with another doesn't change the true value of the expression.

Since the condition requiring two synonyms to be interchangeable in all contexts is too strict, we usually rely on the slightly weaker statement that WordNet synonyms should be interchangeable in a large set of contexts. For example, the word plank replaced by board rarely changes the true value of the expression if one talks about carpentry. However, this replacement is unacceptable in some contexts.

Most synsets are accompanied by explanations similar to those you can find in dictionaries. A word that has several meanings is contained in multiple synsets. The authors consider a synset to be a representation of a lexicalized concept of the English language.

WordNet thesaurus contains words of four parts of speech (nouns, adjectives, verbs, and adverbs), and each part of speech has an individual semantic network. Each part-of-speech synset in WordNet has an individual set of relations. It is assumed that splitting the synsets by a part of speech conforms with psycholinguistic experiments showing that human memory represents information about adjectives, nouns, verbs, and adverbs in different ways.

Note that the latest versions of WordNet include some derivationally related forms that link lexical units of different parts of speech. For example, the verb synset change, alter, modify is related to such nouns as changer, change, alteration, modification and adjectives alterable, modifiable.

Software applications usually use a semantic network of nouns connected using the relations of synonymy, antonymy, hyponymy (hypernymy), meronymy (part-whole), so let's discuss a representation of nouns in detail.

The main relation between noun synsets is a so-called subsumption relationship or "is-a" type relation, in which the subtype synset is called a hyponym, and the supertype synset is called a hypernym. This relation is considered transitive. When B is a hypernym of A, and C is a hypernym of B, it is assumed that C is a hypernym of A. For example, when it is said that a crow is a bird and a bird is an animal, we can conclude that a crow is an animal. The synset X is called a hyponym of the synset Y if native speakers of English accept sentences like 'An X is a (kind of) Y'.

hus, relations between synsets form a hierarchy. When constructing hierarchical systems based on type relations, it is often assumed that the properties of superior concepts are inherited by the inferior ones. It is called the property of inheritance. Thus, nouns in WordNet are arranged into hierarchies as an inheritance system. Systematic efforts have been made to connect hyponyms with their hypernyms for each synset.

The part-whole relations described in WordNet for nouns also include membership relations (member of: for example, a tree is a member of a forest), as well as substance relations (substance of: glass is a substance of glassware).

It is assumed that meronyms can be inherited by hyponyms, for example, if the wing and beak are described as parts of a bird, then all species of birds inherit these parts.

The authors stress that one of the problems of describing meronymy relations is that the parts are described on a higher level than needed. For example, a wheel is often claimed to be a part of a vehicle but then a sleigh is not a vehicle. This is often caused by the fact that a concept on a necessary level is not lexicalized. For this example, a special synset was created in WordNet wheeled vehicle.

Another relation established for nouns is an antonymy relation. An antonymy relation connects two specific words rather than synsets. Moreover, an antonymy relation is not inherited by hyponym synsets. It is assumed that an antonymy relation should be explicitly described. The examples of antonymy relations in WordNet are victory/defeat, happiness/unhappiness, man/woman.

The description of the meanings of polysemous words caused a serious discussion. The average number of meanings in WordNet is greater than in traditional lexicographic dictionaries. Authors of many papers admit that the differences in meanings in WordNet are too subtle for such software applications as machine translation, information retrieval, text classification, question answering systems, etc.

Adjectives and verbs have a particularly large number of meanings. For example, the verb give has 44 meanings and the adjective good has 21 meanings. Some of the meanings go well only with a narrow set of words, for example, the meaning of give19: Give19: give (give (as medicine); “I gave him the drug”).

These issues led to the question of how and what types of meanings of a polysemous word may be united (clustered) in software applications for automatic text processing when the meanings of a polysemous word in the cluster don’t need to be separated from each other, and it will not decrease the quality of the application. Studies show that clustering of meanings can be based on mutually exclusive criteria (semantic, syntactic, and subject-oriented), which also indicates the various significance of different subdivisions of meanings for specific applications in automatic text processing.

4 Calculating Semantic Similarity of Texts Based on Thesaurus

One of the automatic text processing tasks is finding semantic similarity between words. Semantic similarity is often estimated as a value from 0 to 1,

where 1 is the maximum semantic similarity.

According to the basic assumption about calculating semantic similarity based on a thesaurus, the shorter the path between synsets to which the two words in concern belong, the more similar these words are. We also usually focus on the paths in a hierarchy of hyponym-hypernym relations instead of arbitrary paths. This leads us to another important term, namely, Least Common Subsumer. It refers to the closest hypernym found when moving upward the hypernym relations from the two words, for which the semantic similarity is calculated.

These approaches often assume that the path length equal to 1 is a path between synonyms of the same synset. One relation between synsets will be considered as the path length equal to 2.

The approaches to calculating semantic similarity between words based on thesaurus paths include:

- Approaches that use only the length of the path between nodes in a thesaurus (PATH).
- Approaches that use the length of the path between nodes in a thesaurus and the depth of the nodes in a hierarchy.
- Approaches based on information content or self-information.

Path-based measures use only the length of the path between network nodes to calculate similarity. Let's look at one of such measures.

$$\text{path } (a, b) = \frac{1}{\text{shortest \_hypernym \_path } (a, b)}$$

When considering only the length of the path, we lose the information about the concept specificity. The concepts on deeper hierarchical levels are more specific and seem semantically closer to each other than more generic concepts. That's why the measures based on the depth of a hierarchy have been introduced. One of such measures is the Wu and Palmer semantic similarity measure.

$$\text{wup}(a, b) = \frac{2 \cdot \text{depth}(LCS(a, b))}{\text{depth}(a) + \text{depth}(b)}$$

$$\text{depth } (x) = \text{ shortest } \text{ hypernym } - \text{ path } (x, \text{ root })$$

An information content-based measure is defined as the value of the probability to encounter an instance of the concept C in the large corpus $P(C)$. This probabilistic function has the following property. If C_1 is a type for C_2, then $P(C_1) \leq P(C_2)$. The probability value for the top of the hierarchy equals 1. According to the information theory, the information content of the concept C may be represented by the negative logarithm of the probability:

$$\text{IC(C)} = -\log(\text{P(C)})$$

The more abstract a concept is, the less its information content value is. To solve the problem of resolving lexical polysemy, we introduce the concept of Least Common Subsumer (LCS). The similarity between nodes can be calculated by the Lin formula:

$$\text{sim}_{\text{lin}} = \frac{2 \cdot IC\left(\text{LCS}\left(C_1, C_2\right)\right)}{IC\left(C_1\right) + IC\left(C_2\right)}$$

or the Jcn formula:

$$\text{sim}_{jcn} = \frac{1}{IC\left(C_1\right) + IC\left(C_2\right) - 2 \cdot IC\left(LCS\left(C_1, C_2\right)\right)}$$

To find the similarity between words, we can use not only the paths between word synsets but also the structure of the entire graph of the resource relations. One of the methods we can employ is called PageRank. It was initially proposed to rank web pages in search engine results.

PageRank also considers random walks. Assume that a user is browsing random pages. The user starts at some page, then, on every step, the user follows the link from one page to another with an equal probability for all outbound links. Then, in the limit, each page is assigned a view rating, which can be used as a page rank.

However, random walks may end at a dead-end page. To address this, the teleport operation was introduced.

- At a dead-end page, the teleport operation takes the user to another page of the network.
- At pages that have outbound links, the teleport operation takes the user to a random page of the network with some predefined probability (a teleportation coefficient, for example, 0.1), and the remaining probability is used to make the user follow one of the outbound links as before.

PageRank ranks network nodes using an iterative algorithm defined by the formula:

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right),$$

where $(1-d)$ is a teleportation coefficient), $PR()$ is a current rank of the page, $L()$ is a number of outbound links.

Firstly, each network node is assigned a random PageRank value. Then, starting with any page, the PageRank value of the page is recalculated by the formula using the PageRank values of other pages.

Thus, the more the user randomly walks at some page, the more high--ranked links lead to the page and, therefore, the page ranks higher.

To find semantic similarity between words, we can modify the initial PageRank measure as follows. The synset-relation graph is considered a random-walk network. The relations between synsets are considered random-walk

links. To find the words closest to a given word, the teleport operation takes the user only to the synsets that contain the given word as a synonym. Then the remaining synsets and, therefore, words are ordered by their proximity (rank) to a given word.

5 Creating WordNet-Like Thesauri for Other Languages

Creating such massive lexical resources as WordNet from scratch is a challenge requiring many years of guided human labour. To create thesauri for automatic text processing, developers should consistently solve complex lexicographic problems since a system of relations is formalized, main hyponym-hypernym relations are transitive, and the resource has to be designed for the further automatic text processing purposes. Such problems include:

- Splitting a set of synonymous words into synsets.
- Finding ways of representing phrases.
- Extracting systems of meanings of polysemous words.

To accelerate the development of new wordnets in other languages, the first version may be created based on machine translation of Princeton's WordNet into a target language. The obvious disadvantage of this approach is the need to thoroughly review and proofread the generated translations. As an intermediate approach, researchers suggest two-step development of a new language wordnet. The first step includes translating and transferring the relations of the upper 5 thousand concepts of Princeton's WordNet (so-called core WordNet), and the second step is a manual augmentation of hierarchies based on the target language dictionaries and corpora.

Wordnets in other languages preserve the basic structure of the original WordNet. However, they apply different principles of incorporation of words and expressions into synsets, they have another set of semantic relations between synsets, or interpret semantic relations differently. Wordnets may also differ in approaches to polysemy description, which leads to a more fractional or larger system for a representation of meanings of polysemous words. The approaches to the inclusion of phrases into a wordnet may also vary.

The Open Multilingual Wordnet that is being developed aims to link the existing wordnets using open licenses[1]. In 2020, the Open Multilingual Wordnet released 35 wordnets for different languages, and the synsets of each wordnet were linked to those of Princeton's WordNet. All these wordnets are available

[1] http://compling.hss.ntu.edu.sg/omw/

as individual files in several formats. The wordnets are accessible through the (python) Natural Language Tool-Kit wordnet interface, or NLTK[2]. To link a wordnet in a new language, it's necessary to link the synsets of the new language to those of Princeton's WordNet and provide data in the required format.

6 Wikipedia as a Multilingual Ontological Resource

Wikipedia is known as an online encyclopedia created and maintained in many languages by a community of volunteer editors. Wikipedia can be considered a crowdsourcing project involving many people around the world. At the same time, Wikipedia is a free encyclopedia of so-called free content defined as content that does not bear copyright restrictions on the right to redistribute, study, modify and improve, or otherwise use works for any purpose.

Thus, the knowledge gained from Wikipedia and applied to different tasks including automatic text processing applications imposes no charges on developers. Since the launch of Wikipedia, many papers have used its content to solve different problems of automatic text processing, including automatic text classification, calculating the semantic similarity of texts, finding similarities between search queries, extracting keywords, etc. Wikipedia-based approaches may use a system of relations between Wikipedia pages and a concept representation in the form of a text vector based on the corresponding article of Wikipedia.

We can also think of Wikipedia as of a multilingual ontological computational resource that can be used for automatic text processing. Each Wikipedia page provides information about a specific concept or named entity. There are several types of relations between Wikipedia concepts:

- Redirection relations link specific linguistic units and Wikipedia pages to each other, and each concept can be associated with several words and expressions. Thus, words and expressions linked to the same concept form a list of synonyms, a synset similar to a synset of WordNet-like thesauri.

- Meaning relations (disambiguation pages) describe polysemous linguistic units that lead to several Wikipedia pages. It is also similar to the description of polysemy in wordnets, in which polysemous words are incorporated into different synsets.

- Internal relations between concepts are represented by hyperlinks in a text. They lead to related pages, creating a system of relations between the Wikipedia concepts.

[2] http://www.nltk.org/howto/wordnet.html

- Multilingual relations that exist between Wikipedia pages describe the same concept in different languages.

- Categories are Wikipedia pages linked to one or several categories of two types: set categories, such as C, Cities, this category includes pages about cities (New York, Paris, etc.), and topic categories, for example, C-City, this category consists of pages about city-related information (city planning, urbanization, etc.).

Thus, Wikipedia can be considered a formalized representation of knowledge about the world, i.e., an ontological resource. We can look at it as a graph in which vertices or nodes are Wiki pages and edges are the relations between Wiki pages. We can also represent WordNet-like resources as a graph in which vertices are synsets, and edges are the described semantic relations.

Such representations of WordNet and Wikipedia allow setting the problem of combining WordNet and Wikipedia into a single ontological resource with wider coverage of existing concepts and instances than in WordNet. At the same time, WordNet could introduce semantic relations described in a more strict fashion into the Wikipedia structure.

Combining linguistic resources (such as WordNet-like thesauri) and a semi-structured resource (for example, Wikipedia) makes it possible to create a multilingual resource in the form of a semantic network. BabelNet is one of such well-known resources.

To link WordNet and English Wikipedia, BabelNet uses the following information about the original resources:

- All available synsets and word meanings in the current version of WordNet 3.0, as well as their lexical and semantic relations.

- All available rich-in-content pages of Wikipedia that are considered Wikipedia concepts, as well as associative relations between them, extracted based on hyperlinks between the corresponding pages.

To create one multilingual resource, it's necessary to:

- Merge the corresponding WordNet synsets and Wikipedia concepts into the so-called BabelNet synsets.

- Enlarge multilingual text inputs of BabelNet synsets using a) the existing Wikipedia links to pages describing the same concept in different languages and b) an automatic translation system.

- Connect BabelNet synsets using all WordNet relations, as well as all associative relations of the corresponding Wikipedia concepts, while extracting relations from Wiki pages in all languages that are being processed in the current version of BabelNet.

BabelNet consolidates such multilingual resources as the Wiktionary and wordnets of the Open Multilingual Wordnet, that's why BabelNet includes words and expressions in more than 270 languages. BabelNet improved the quality of resolving polysemy in multiple languages, as well as that of a related entity linking problem, i.e., linking the entities mentioned in the text to a base resource, such as Wikipedia.

7 Conclusion

In general, lexical and semantic resources (thesauri, linguistic ontologies, and semantic frames) are highly demanded in automatic text processing, and resource integration increases the coverage and multilingualism of such resources.

Lexical and semantic resources are used in such applications as:

- Text semantic analysis.
- Semantic conceptual indexing in information retrieval and analytical information systems.
- Development of additional purpose-specific resources, for example, ImageNet and SentiWordNet.
- Enrichment of text embeddings that integrate knowledge into statistical methods (probabilistic topic models and distributed word embeddings). Word embeddings where words are represented as vectors often suffer from such problems as meaning collision in one vector and various difficulties with a representation of phrases. Moreover, very close synonyms or different names of the same object may have different embeddings.
- As a source of features for machine learning systems, including deep learning approaches and so on.