Natural language processing is the study of computer programs that can understand and produce human language. An important goal in the research to produce such technology is identifying the right meaning of words and phrases. In this paper, we give an overview of current research in three areas: (i) inducing word meaning; (ii) distinguishing different meanings of words used in context; and (iii) determining when the meaning of a phrase cannot straightforwardly be obtained from its parts. Manual construction of resources is labour intensive and costly and furthermore may not reflect the meanings that are useful for the task or data at hand. For this reason, we focus particularly on systems that use samples of language data to learn about meanings, rather than examples annotated by humans.

Keywords: natural language processing; word meanings; semantics; computational linguistics

1. Introduction

Natural language processing (NLP) research studies computer programs that can understand and produce human language both from a theoretical standpoint and for practical purposes. The ability to converse with machines using one’s own language has excited much interest owing to the obvious benefits of avoiding learning a specialized computer language or interface. Human language is extremely complex and while we do not yet have fully conversant machines, new horizons are emerging that NLP technologies might help us with. The World Wide Web contains a wealth of information for our personal and business lives, much of it in language form. While we can obtain information from keyword search, it would be useful to get pertinent summaries to our questions, especially since information might need collating from several documents and the documents may not be written in the languages that we speak. The potential benefits from systems that can translate, understand and summarize language data are proliferating as the volume of information in language form on the World Wide Web increases.

Relatively simple statistics from large samples of text are useful in many tasks (Keller & Lapata 2003). For example, frequencies of short sequences of two to five words can help predict words in speech recognition. Statistics collected from previous examples can predict the likelihood of a translation for a given sentence. However, many researchers believe that some deeper representation of meaning is required for a computer to manipulate language as we do. Words have meanings.

* dianam@sussex.ac.uk

One contribution of 20 to a Triennial Issue ‘Chemistry and engineering’.
Humans learn what a word represents and when it can be used. If computers are going to answer our questions, make non-trivial paraphrases and summaries, translate between languages and produce documents or ask questions themselves, then they will need to know the meaning of words. In §2 of this paper, we will discuss some current approaches that have been used for representing word meaning in NLP systems. Many systems use electronic dictionaries and thesauruses handcrafted by people. We contrast these with ways in which meaning can be induced automatically by using large volumes of language data and using information about the distributions of words to infer meaning.

The representation of word meaning will be impacted by the fact that many words have more than one meaning. For example *star* can mean: (i) celestial body, (ii) celebrity, (iii) shape, and (iv) a zodiac sign. While many words in a dictionary may be listed with only one meaning, the more common words that we use in everyday language tend to have more meanings. We do not note all the meanings (also referred to as senses) because one meaning is usually made obvious from the context. Computers need a mechanism whereby the meanings of words can be determined automatically from the context. This is referred to as word sense disambiguation (WSD). There has been much work in this area using manually produced dictionaries and thesauruses as inventories of word meaning. Automatically induced inventories are seldom used for disambiguation. There are two main reasons for this. Firstly, the standard datasets that have been used by the community for evaluating the techniques are typically samples of text labelled with a predefined inventory of meanings from a manually created dictionary or thesaurus. In order to compare an automatically induced inventory to the state of the art, a mapping is needed between the induced and the predefined inventories. This puts automatic inventories at a disadvantage. Secondly, the systems that perform best at the task use not only the predefined listing of senses but also hand-labelled examples to learn which sense goes with which context. Systems that do not need hand-labelled examples do not perform as well; however, research in this direction is vital because the hand-labelling is costly and there are always words, domains and languages without such data. In §3, we will provide a summary of work on WSD. We discuss issues that have been raised concerning the standard evaluation methodology and point to new directions that will allow the evaluation of fully automatic systems for disambiguation.

The meanings of words on their own are not alone sufficient for systems that need to understand, translate and summarize language. Systems need to understand the meaning of phrases and sentences as well as longer units of discourse. This paper focuses on meanings of individual words and does not cover material on the meanings of regular combinations of words into larger units. However, we introduce work that is being done to detect idiomatic usages where a particular combination of words is used to mean something different from its literal interpretation. In English, for example, *get the drift* is typically used in a context that has nothing to do with snow or something moving on water or air. There have been a number of proposals to find combinations of words that are not always used literally. It is crucial for computers to be able to deal with such idiomatic usages so that they do not

---

1 In this paper we use **bold face** when describing meanings.
make wrong interpretations. One would not want documents with *I will eat my hat* retrieved if one was trying to compile a list of diets. We will give an overview of research in this direction in §4 before concluding the paper in §5.

2. Representing word meanings in a computer

We divide this section into two sub-sections. Section 2a describes a couple of handcrafted resources that are frequently used in NLP for representing meaning. Most of these resources were not specifically designed for NLP but have been adopted by researchers owing to the useful information they contain. Since it is the potential for application, rather than their construction, which interests researchers we will not cover manually produced resources in any breadth or depth. Section 2b describes work on automatically inducing information about word meaning from language data.

(a) Manual inventories of word meaning

A good deal of research on word meaning focuses on the disambiguation process using man-made dictionaries or thesauruses to represent meanings. Many researchers use dictionaries such as the Longman Dictionary of Contemporary English (LDOCE; Procter 1978) owing to the wealth of information it contains. For example, LDOCE contains sub-groupings of meanings, and preferences that words in specific grammatical relationships have for each other. For example, the verb *eat* has a preference for *food* as an object. It also lists the meanings of a word according to how prevalent they are in everyday use, which we will see below is very important information indeed. This information has proved very useful for WSD (Stevenson & Wilks 2001).

*WordNet* (Fellbaum 1998) in contrast is a thesaurus based on psycholinguistic principles which represents the meaning of words by the relationships the words have with other words. It is widely used because it is publicly available, is full of useful information and has been used in the construction of many datasets for disambiguation evaluation. In *WordNet*, words are grouped together into ‘synsets’ that contain near synonyms (words that mean nearly the same thing), for example *car, auto, automobile, motorcar*. The synsets are related to links such as ‘is-a’ links (hyponymy), for example *apple ‘is-a’ fruit*; figure 1 shows a graphical example with a small sample of words and relationships.

(b) Automatically induced inventories

While inventories of word meaning produced by humans have been very useful for NLP research to date, researchers are looking into ways of inducing word meanings from language data. These methods rely on large samples of language data, called corpora (corpus in the singular). Such corpora are now plentiful due to the work of corpus linguists who have created these for a variety of linguistic studies (Sampson & McCarthy 2004), for example the British National Corpus (BNC; Leech 1992) that contains 90 million words of English text and 10 million words of spoken English. There is also now a great deal of interest in obtaining corpus data from the web (Keller & Lapata 2003; Brants & Franz 2006; Sharoff 2006), since it contains so much data and that too in a variety of languages and domains.
Most automatic inductions of word meaning inventories use the distributional hypothesis (Harris 1968) that words that occur in similar contexts tend to mean similar things. The methods collect frequency counts of the contexts that the target words occur in. Contexts are typical words occurring within a fixed window of words on either side of the target. Often only content words are used, i.e. nouns, verbs adjectives and adverbs, but not grammatical words such as prepositions (e.g. of) or determiners (e.g. the). A sentence such as

\[ \text{We have also proposed that they continue to encourage ways of giving greater priority to coaches on the national road network following the successful introduction of a dedicated lane on the M4 Spur to Heathrow.} \]

taken from the English Internet Corpus (Sharoff 2006) might yield \textit{encourage, way, give, great, priority, national, road, network, follow, successful} as the contexts in a 10-word window around the target word \textit{coach}.\(^2\) Table 1 provides a made-up example for five target words given in four contexts: \textit{take, teach, team} and \textit{drive}.\(^3\) Typically, several hundred contexts are chosen automatically by using a procedure that finds the most informative contexts from all possibilities (Schütze 1998). The words are represented as vectors with an index for each context and the value at that index is the frequency. Figure 2 gives a visual

\(\text{Table 1. Frequency data input to a vector space model.}\)

<table>
<thead>
<tr>
<th>target words</th>
<th>context</th>
<th>coach</th>
<th>bus</th>
<th>carriage</th>
<th>trainer</th>
<th>instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>take</td>
<td>50</td>
<td>56</td>
<td>25</td>
<td>20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>teach</td>
<td>30</td>
<td>10</td>
<td>4</td>
<td>29</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>team</td>
<td>35</td>
<td>3</td>
<td>0</td>
<td>20</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>ticket</td>
<td>22</td>
<td>15</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Most automatic inductions of word meaning inventories use the distributional hypothesis (Harris 1968) that words that occur in similar contexts tend to mean similar things. The methods collect frequency counts of the contexts that the target words occur in. Contexts are typical words occurring within a fixed window of words on either side of the target. Often only content words are used, i.e. nouns, verbs adjectives and adverbs, but not grammatical words such as prepositions (e.g. of) or determiners (e.g. the). A sentence such as

\[ \text{We have also proposed that they continue to encourage ways of giving greater priority to coaches on the national road network following the successful introduction of a dedicated lane on the M4 Spur to Heathrow.} \]

taken from the English Internet Corpus (Sharoff 2006) might yield \textit{encourage, way, give, great, priority, national, road, network, follow, successful} as the contexts in a 10-word window around the target word \textit{coach}.\(^2\) Table 1 provides a made-up example for five target words given in four contexts: \textit{take, teach, team} and \textit{drive}.\(^3\) Typically, several hundred contexts are chosen automatically by using a procedure that finds the most informative contexts from all possibilities (Schütze 1998). The words are represented as vectors with an index for each context and the value at that index is the frequency. Figure 2 gives a visual

\(\text{Figure 1. A small portion of WORDNET.}\)

\(\text{Figure 2. A visual representation of the data in Table 1.}\)

\(^2\) Note that is common practice to stem words to their root form so that \textit{ways} becomes \textit{way}, although this is not always done.

\(^3\) This example just shows words from the same part-of-speech category (noun), but one could look at similarities of words from different parts-of-speech.
illustration of how the words in table 1 might look using only two dimensions.\footnote{We only use two dimensions for illustration.} Using several hundred dimensions will partition a large variety of words. Words that have similar meanings tend to occur together in vector space. Many words, of course, have several meanings. Coach can mean bus or instructor; thus words such as this tend to occur somewhere between the different meanings, exact positioning depending on the frequency of occurrence of the meanings in the corpus. Schütze (1998) has worked on determining meanings of a word by clustering example sentences using the vector representation of the other words in the sentence. So, for example, the sentences with coach might be clustered into two groups as s1 and s2 in figure 2 reflecting the different meanings of coach.

There are other approaches that use distributional data for finding word meanings. One common approach is to compare the contexts of a target word directly with others (potentially all other targets) using a measure that estimates the distributional similarity of one word with another, i.e. how much commonality in the contexts of occurrence is there between the two words. There are many different distributional similarity scores (e.g. Lin 1998; Lee 1999; Weeds & Weir 2003). The output of this approach is a thesaurus where each word is listed with a ranked list of the \( k \) most similar words (referred to as \( k \) nearest neighbours). The distributional similarity score between the target word and the neighbour is provided. As an example, here is the output for the noun coach in a thesaurus created from the BNC using Lin’s distributional measure (Lin 1998) to get the 10 nearest neighbours:

\begin{verbatim}
coach: train 0.171, bus 0.166, player 0.149, captain 0.131, car 0.131, club 0.131, teacher 0.129, team 0.129, boss 0.128, chief 0.128.
\end{verbatim}

The distributional similarity scores specify just how close the neighbours are.

As with the vector space models, there is some interesting work on teasing out different meanings from the target word with these \( k \) nearest neighbour approaches. Pantel & Lin (2002) cluster the neighbours to find subgroups with common contexts. Thus, the coach neighbours might be transformed to two groups, for example (train, bus, car) and (player, captain, club, teacher, team, boss, chief).

While there are errors and inconsistencies in manually defined resources, these automatic methods also have their own problems. The input data will contain errors such as spelling mistakes and there will also be errors introduced by
automatic procedures, for example procedures that split the data into sentences or decide whether a word like *fly* is functioning as a noun or verb in a given sentence. Additionally, there are issues because the distributional approach will return words that have a variety of relationships with the target word, not just synonymy. Some words have near synonyms, for example *car* and *automobile*, whereas others do not, for example *apple*. Regardless of whether a candidate has a synonym, often more specific or more general, semantically related words will be returned, for example *pear fruit food* might be returned for *apple*. Furthermore, distributionally similar words can have opposite meanings, for example *hot* and *cold*, or *white* and *black*, because they occur in similar contexts. Research has been done to find out how distributional measures are affected by frequency (*Weeds et al.* 2004) and whether the contextual features used can distinguish different relationships (*Padó & Lapata* 2003). While these automatic approaches are not without their problems, they open up exciting prospects in computational understanding because the output is determined by language data rather than being subject to the decision making of dictionary producers (*Kilgarriff* 2006).

For a computer system to correctly interpret language, it needs to not only identify different meanings but should also be able to determine which meaning is being used in a particular context and it is to this topic that we now turn out attention.

### 3. Disambiguating word meanings

Ambiguity is pervasive in natural language although we are often not conscious of it because contextual clues from the dialogue or environment determine the meaning. If a computer system needs to operate in applications where words are going to take different meanings, then we need to build a mechanism for disambiguation into our computers. The majority of computer systems designed to do this task fall into two categories. The first category consists of supervised approaches that exploit hand-labelled examples so that computers can use the examples as training data to learn associations between contextual clues and senses. For an overview of such methods see *Márquez et al.* (2006). The second category comprises knowledge-based approaches that use information contained in manually produced dictionaries and thesauruses. For example, information such as the hyponyms and textual definitions of synsets from *WordNet* (described in §2) can be used to find the most probable meaning of a word given the words that co-occur in the context. For an overview of these methods see *Mihaic* (2006). There are also hybrid approaches that use both hand-labelled data and handcrafted knowledge. Many of the systems also supplement the hand-labelled data or handcrafted knowledge with frequency information gleaned from unlabelled corpus data. There is a third category of computer systems that perform WSD using inventories automatically induced from corpora (see *Pedersen* 2006). These have not been widely used for WSD for reasons that we expand upon below.

There has been a massive amount of research on WSD. This has proliferated due to a series of international evaluation exercises known as Senseval, which has created standard datasets for this task. At the time of writing, there have been three such evaluation exercises, and a successor (renamed SemEval) is taking place this year. In these exercises, computer systems compete on various WSD tasks.
The scenario is as follows. A sample of text is selected and human annotators are given a predefined inventory, such as WordNet. They then have to label the target words in the text with whichever sense fits the context best. Sometimes the task focuses on a specific sample of words and there are many sentences from a variety of different documents for each word. Another approach is to label all content words in a document. The computer systems automatically do the same task, given the inventory and are scored against the verdicts of the human annotators. The datasets have been a major catalyst to the field, but two major issues have arisen.

Firstly, systems that use hand-labelled data to train do best, but often only a little better than a heuristic of taking the first listed sense in the inventory (Snyder & Palmer 2004). Hand-labelling of data with senses is not a sustainable method because it is very costly and there will always be words, domains and languages without sufficient hand-labelled data. Even using the heuristic of the first sense can be problematic as the frequency of senses will depend on the data one is using.

Secondly, there has been a great deal of debate as to which inventory is best for WSD. The distinctions in manually defined inventories depend on the decisions of lexicographers and dictionary producers and do not necessarily either have cognitive validity or meet the needs of an NLP application (Kilgarriff 2006). WordNet has been widely used as it is full of useful information and is freely available for research purposes. It is a truly remarkable resource, but there is concern that the meaning distinctions are too fine-grained for applications (Ide & Wilks 2006); for example, there are 59 meanings of the verb break. Grouping senses to coarser distinctions may help, but it is not always apparent which groupings are needed. For example, in Senseval-2 groupings were provided for child, which has four meanings in WordNet. Young person was grouped with immature person, while offspring was grouped with member of a clan. While this youth versus descendant distinction may make intuitive sense, it leads to a loss of relationship between young person and offspring which are often translated to the same word, for example, niño/a in Spanish. There is also a problem that any predefined inventory will make it difficult to compare systems with a different inventory. The results will be masked by errors in mapping to the predefined inventory used for annotation.

In §3a, b, we point to recent research that attempts to address these two issues.

\(a\) Automatically determining the most probable meaning of a word without hand-labelled data

There is a tendency for one meaning to be predominant in a given text. Most WSD systems use hand-labelled data to determine the most probable meaning and if the right meaning is not evident from the context, they just pick the predominant meaning, which gives them the right answer most of the time. However, there are many languages without hand-labelled data, furthermore the meaning of a word depends on the domain of a document. For example, star in a popular press article is more likely to mean celebrity, whereas in a scientific article it is more likely to mean celestial body. Even in general text, there is not enough hand-labelled data available to provide the most probable meaning. For example, given the publicly available hand-tagged corpus for WordNet (SemCor; Miller et al. 1993) the most frequent meaning of tiger is audacious person rather than animal because there are too few examples of tiger in the 200 000 word sample.
Researchers at the University of Sussex developed a technique to automatically find the most probable meaning of a word given a sample of text (McCarthy et al. 2004). The method has been tested on Senseval and SemCor datasets, as well as datasets created from hand-labelling data from different domains (Koeling et al. 2005). The method uses distributional similarity to produce a list of nearest neighbours for a target word, as in §2. The neighbours are automatically linked to the meanings in the sense inventory (WordNet) using measures that use information from the inventory itself.

Of course, detecting the most probable meaning of the word does not solve the problem. There are many cases where the dominant meaning is not used and the local contextual clues are needed. Further work is needed to find local contextual clues automatically, for example, by using the contexts that the neighbours have in common with the target word. There is also promising research into automatic acquisition of domain-specific information (Koeling et al. 2007).

(b) Evaluating disambiguation techniques without specifying a sense inventory

There is no consensus on the right inventory of word meanings to use for a given application, with the exception of machine translation that can exploit alignments of data between the source and target language (Resnik & Yarowsky 2000). Generalizations to other tasks are problematic because the distinctions depend on the languages chosen, and if many languages are considered then the distinctions become very subtle (Palmer et al. 2007).

It has not been possible to compare automatic induction of word meanings on an equal footing because datasets use predefined resources. In SemEval this year, a new dataset (McCarthy & Navigli 2007) has been produced which aims to address this issue. For this task, annotators and systems provide a substitute word for a target word in context. For example, a substitute of game might be given for the word match in the following sentence:

After the match, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament.

For any test item, there will be many responses that could be appropriate, so measures that reflect agreement between annotators will be lower than on a task where annotators are selecting an item from a fixed inventory, however the inventories that people use are as much a part of the task as the disambiguation process. By using substitutes, the distinctions made can potentially have an impact on tasks like text simplification where a system needs to reword sentences to meet the needs of a client group (Devlin & Tait 1998) or perform question answering where the answer to a user’s question may be worded differently in the database or document collection. In figure 3a, we show the synonyms used by annotators for each of the 10 sentences in the test data with the word fire. Here the synonyms show two core meanings sack and shoot, although there is also evidence of a subtler distinction between shoot the person and shoot a gun where the former has the synonym shoot at and the latter has the synonyms discharge and launch. Figure 3b shows the data for investigator where we see overlap between the meanings researcher and detective. This reflects the relatedness of different meanings. We look forward to comparisons of all inventories, man-made and automatically induced, on this dataset.
The meanings of words are determined by their contexts and often even a phrase provides sufficient context if a computer system has the necessary information to tie the phrase to the meaning. However, another issue that computer systems must deal with is the distinction between literal and idiomatic usages, and it is to this issue that we now turn our attention.

4. Words with spaces

While words, or their base forms, are often thought of as the basic unit of meaning, there is increasing interest in phrases that function as a unit rather than as a straightforward composition of units. These ‘words with spaces’ or ‘multiwords’ exhibit a spectrum of idiosyncratic behaviour (Sag et al. 2002). The lack of transparency may exist to a greater or lesser extent. So for example, while traffic light means a specific sort of traffic signal with a vertical array of one red, one orange and one green light, a non-native speaker might easily pick up the phrase very quickly without being explicitly told, whereas I will eat my hat might need an explanation.

Figure 3. Synonym distributions from annotators. (a) Fire (verb) and (b) investigator (noun).
NLP systems have to deal with such phrases in order to interpret language appropriately. Some work has been done using distributional approaches for determining word meaning to ascertain when a particular combination has an idiosyncratic interpretation (Baldwin et al. 2003; Bannard et al. 2003; McCarthy et al. 2003) by contrasting constituent words with the phrase to see if the meaning is transparent or not. Baldwin et al. (2003) test their methods on noun–noun compounds, for example traffic light, and phrasal verbs, for example blow up, while Bannard et al. and McCarthy et al. test their methods only on phrasal verbs. If the phrase is not close in vector space to the constituent words (Baldwin et al. 2003; Bannard et al. 2003) or if there is little overlap in the nearest neighbours of the phrasal verb blow up and the verb blow (McCarthy et al. 2003), then the phrasal verb is likely to be more idiomatic.

Another approach (McCarthy et al. 2007) uses distributional methods to find classes of the sorts of words that occur in a grammatical relationship with a target, for example, words related to food tend to occur as the objects of eat. When a particular combination seems at odds with the anticipated type of object, for example eat hat, this indicates that the combination might be idiomatic.

Most research has focused on detecting whether a given candidate is idiomatic. There has been subsequent work that uses vector space models to demonstrate that we can use local context to automatically distinguish between literal and idiomatic usages of the same phrase (Katz & Giesbrecht 2006).

More work is needed both to find phrases that are sometimes used idiomatically and to distinguish idiomatic usages from literal ones just as we distinguish word meanings. Currently, many systems cannot do this as was demonstrated on the SemEval substitution task described in the previous section. The annotators were asked to indicate if the word was an integral part of a phrase, and also they were permitted to supply phrasal substitutes if they could not think of a reasonable single word substitute. The participating systems performed much better on the data that did not involve such ‘words with spaces’. Most systems do not handle idiomatic phrases and so research is needed in this important area to prevent our computers from getting the wrong end of the stick.

5. Conclusions and future directions

In this article, we have focused on the meanings of words and idiomatic phrases. Work on automatic induction of meaning is an exciting area of research that could help us achieve our goal of building systems that can understand and produce language. While a lot of research has exploited man-made online resources, there are greater benefits in using methods that rely on language data rather than human effort because human effort is costly and resources are not always available. Currently, such methods perform less well than those that use manually produced resources, but they show potential given the increasing amount of data available on the World Wide Web. So far the data used are typically in written form, though learning from spoken language is also possible. One exciting area that is yet to be fully exploited is coupling text or speech with visual clues (Barnard et al. 2003; Gold & Scassellati 2007; Liu et al. 2007). We look forward to the use of these methods and many more of the kind in our endeavour to enable computers to get the drift.
References


Phil. Trans. R. Soc. A (2007)


Phil. Trans. R. Soc. A (2007)
Diana McCarthy is a Royal Society Dorothy Hodgkin fellow in the Department of Informatics at the University of Sussex. Her research interests in natural language processing primarily concern automatic lexical acquisition. Diana was born in Pembury, Kent in 1964. She was awarded her first degree in Speech Pathology by the University of Manchester in 1987. She practised as a speech and language therapist for several years before undertaking a Masters degree at the University of Sussex in knowledge-based systems out of a curiosity as to how one might use a computer to model human reasoning, and language in particular. After being awarded her Masters in 1989, she worked in the commercial sector applying artificial intelligence techniques to meet customer needs before returning to academia in 1995. She obtained her PhD from the University of Sussex in 2001 and worked on a number of EU and EPSRC projects on lexical acquisition and disambiguation. In 2004, she was awarded the best paper prize at the Annual Meeting of the Association of Computational Linguistics for her work with colleagues at the University of Sussex on finding the most probable meaning of a word from a sample of text. She was awarded the best poster prize for a work presented at the Women in Computing: Grace Hopper Colloquium held in May 2007 by the British Computer Society. She is fascinated by computational approaches that learn about words and their meanings from data without human instruction and has recently co-organized an international competition aimed at comparing such approaches with those that require human supervision.