Natural Language Processing & Applications

Introduction

1 Why study Natural Language Processing (NLP)?

Language is the defining characteristic of human beings. Although other primates, particularly chimpanzees and bonobos, can be taught to make some limited use of language, their abilities are far removed from those of even quite young children. Only very severely handicapped infants fail to learn language with little or no direct teaching. Hence there are two powerful and inter-connected reasons for a student of computing to learn something about NLP:

- If computers are to fit in with us, rather than we with them, they will have to be programmed with the ability to handle natural language. Compared to speaking and listening, keyboards and pointing devices are clumsy and tedious. Considerable progress has been made in recent years, and strong commercial pressures will ensure that this progress continues. NLP is likely to become ever more important and hence a professional computer scientist should know something about it.
- Since the very nature of human beings is tied to language use, studying artificial NLP gives valuable insights into human thought.

In deciding what to teach in this module, I had three general aims:

- to give a broad introduction to the topic
- to introduce the minimum necessary knowledge of linguistics
- to show some applications in enough detail for you to be able to assess their achievements and limitations.

I have not worried too much about being comprehensive nor about covering the 'latest ideas'.

Lectures and handouts are biassed towards coverage of the underlying linguistic theory; once you understand this, it's relatively easy to read up on the computing aspects of NLP.

2 Natural Languages

We can define a **natural** language (NL) as one of the languages which humans naturally speak (i.e. excluding human-invented languages such as Esperanto). The term **language** is more difficult to define. Probably no two people use exactly the same set of words with exactly the same pronunciation, grammar and meaning (indeed neither does one individual from one occasion to the next). Hence we can start with an **idiolect**: the language that one individual uses. The next level up is generally called a **dialect**: a set of very similar idiolects. When dialects become sufficiently different to be called languages is generally a political rather than a linguistic question. The division of Serbo-Croat, the common language of former Yugoslavia, into two languages, Serbian and Croatian, shows this rather sharply. Further examples of very similar languages which might be called dialects are: Dutch (spoken in the Netherlands) and Flemish (spoken in north-western Belgium); and Hindi and Urdu (spoken in northern India and Pakistan, which are associated both with political divisions and with the Hindu and Muslim religions respectively). On the other hand, languages used in China which are mutually un-intelligible when spoken are often called dialects of one Chinese language because they use the same writing system.

Estimates of the number of languages spoken in the world vary, but most are in the range 4,000 to 7,000. However, many of these have few speakers and are on their way to extinction (according to UNESCO in 2007, a language disappears disappears on average every two weeks). About half the world's population speaks one of 8 languages. In order of the number of speakers, these are Mandarin (Chinese), English, Hindi, Spanish (Castillian), Russian, Arabic, Bengali and Portuguese.

Linguists divide languages into a number of families, based on similarity and shared descent. Interestingly, of these 8 languages, all but Mandarin and Arabic belong to the same family,

Indo-European, whose influence seems to be increasing. Indo-European languages were natively spoken in a broad band through Europe and on through south Asia to Bengal. Only Finnish, Hungarian, Basque and Turkish intruded into this area. It is believed that all the Indo-European languages are descended from one language spoken around 4,000 BC. In this module, I shall mostly discuss English, but with some examples from other, usually Indo-European languages. The Indo-European family is divided into several subfamilies, including:

- Germanic: e.g. English, German, Norwegian.
- Italic/Romance: e.g. Spanish, French, Romanian.
- Slavic: e.g. Russian, Polish, Czech, Bulgarian.
- Indo-Iranian: e.g. Hindi, Bengali, Persian.

Although Indo-European languages are widely spoken, it is estimated that they make up no more than 3% of the world's known languages. It is important to be aware that different language families may be based on quite different principles, both in their sounds and in their grammar.

3 Levels

Natural languages can be considered to have three very broad 'levels': sound systems, grammar and meaning. To a large extent these levels can be considered to be independent of one another.

Different languages clearly have different **sound systems**, i.e. they use different sounds and combine them in different ways. For example, the *th* sound in *think*[1] does not occur in French; the *eu* sound in the French word *peur* (= fear) does not occur in English. Although the *ng* sound in *sing* occurs in English, it never does so at the beginning of a word, with the result that English speakers find it difficult to pronounce words which start with this sound (e.g. the name of the former Ghanaian president, *Nkrumah*). The study of sounds will later be divided into **phonetics** and **phonology** (roughly the study of the actual sounds and the study of 'distinctive' sounds).

Equally clearly, different languages have different **grammars** – rules governing the formation of words (**morphology**) and of sentences (**syntax**). Consider these English and German sentences:

I eat fish.	*Ich esse fisch.*

The syntax (ordering of the words in the sentence) appears to be the same. However, the underlying morphological rules are different. If we change *I* to *he* (and correspondingly *ich* to *er*), *eat* adds *s*, whereas *esse* changes the terminal *e* to *t* (and, in this particular word, the initial *e* to *i*):

He eats fish.	*Er isst fisch.*

In other languages both syntax and morphology can be very different. Consider Japanese versions of the same two sentences:

Watashi wa sakana o tabemasu.	*Kare wa sakana o tabemasu.*
I subject-marker fish object-marker eat	He subject-marker fish object-marker eats

Meaning is the most difficult of the three levels to define. **Semantics** refers to the meaning of words and of sentences viewed as combinations of words. Do the English, German and Japanese sentences have the same meaning? Consider the verb in the sentences beginning with 'I'. Although the ACTION implied by the words *eat*, *esse* and *tabemasu* may be the same, in several respects these words do not have the same meaning. Word meanings involve both parallels and contrasts. Thus the meaning of *eat* involves both its similarity to and its difference from the word *drink*. Now in English, *eat* contrasts with *am eating*; in German *esse* covers both forms. Thus the German sentence has a broader meaning than the English sentence. In Japanese, *tabemasu* contrasts with *taberu*; both mean 'eat', but the first word is used in formal speech, the second in informal speech. Thus the Japanese sentence conveys a

---

[1] The convention which I shall use in the handouts for this module is that when letters, words or sentences are entities under discussion, they will be in *this format*. For emphasis, THIS FORMAT will be used.

shade of meaning that is not (cannot?) be conveyed in either the English or German sentences. Both *watashi wa sakana o tabemasu* and *watashi wa sakana o taberu* would have to be translated as *I eat fish*. Yet the difference is very important in Japanese. Notice that the Japanese verb remains completely unchanged when 'I' is changed to 'he': formality vs. informality is marked in all Japanese verbs, whereas person and number are not. Examples like this raise very important questions about the possibility of exact translation between one NL and another.

People use language for different purposes, and these purposes also form part of meaning. **Pragmatics** refers to the study of the purposes and intentions of the language user. The syntactic and semantic form of a sentence can disguise its pragmatic purpose. *Can you swim?* is a question, usually asking for information, so that *Yes* or *No* are valid answers. *Can you pass the salt?* appears similar in terms of its syntactic structure and of the meanings of its individual words, but is normally not a question but rather a request for action, so that a *Yes* answer unaccompanied by the action is not a valid response.

The following scenario tries to clarify the levels discussed above. Inevitably it makes use of material to be covered later, so you will benefit from re-reading it at the end of this module. Suppose that two people are alone in a room. One says to the other *John inputs the data*. This utterance can be analysed at six conceptually distinct levels.

1. Pragmatic. The speaker uttered the sentence for a reason. It may be purely declarative or informative, intended to tell the listener a fact he or she did not know. The syntax and semantics of the sentence suggest this. However, it could be a disguised instruction. Thus if the listener is currently inputting data, *JOHN inputs the data* can be used by the speaker to suggest to the listener that he or she should stop and allow John to do it.
2. Semantic. Words and sentences have meanings. *John* refers to a specific individual known to the speaker and listener. Here, to *input* is to enter information into a computer. The order of the words in the English sentence shows that John is the 'Agent', i.e. is doing the inputting. Combining a knowledge of word meanings and a knowledge of the rules governing the meaning of word combinations and orderings will generate a semantic representation for this sentence.
3. Syntactic. Words are systematically ordered to form sentences. The syntactic structure of this English sentence is something like $[John_{\text{noun}}]_{\text{NP}}$ $[inputs_{\text{verb}}$ $(the_{\text{det}}$ $data_{\text{noun}})_{\text{NP}}]_{\text{VP}}$. If the listener replies *Oh, I thought Jane does it*, the words *does it* refer back to the component bracketed as $[..]_{\text{VP}}$ above, i.e. in this example refer to *inputs the data*. If the listener misheard the last word, he or she could say *John inputs WHAT?* or *What does John input?* The equivalence of these two sentences is part of the syntax of English. Syntax can be described by rules.
4. Morphological or morphemic. Words can have an internal structure. The word *inputs* consists of three distinct parts: the prefix *in*, connected with the word *in*; the root *put*; and the 'third person singular marker', spelt *s*.
5. Phonological or phonemic. Words consist of a sequence of DISTINCTIVE sounds or phonemes. As English spelling is notoriously only partially phonemic, *John inputs the data* does not show these very clearly; e.g. the *h* in *John* is silent and the two *a*s in *data* correspond to very different sounds. Using IPA (the International Phonetic Alphabet) enables a more accurate representation of the phonemes: /ʤɒn ɪpʊtz ðə deɪtə/.
6. Phonetic. If the speaker uses 'Standard English English', he or she will most likely produce a sequence of sounds which can be written as [ˈʤɒnˈɪmpʰʊtsðəˈdeɪtə]. There will be no pauses between the words, and some of the sounds will change slightly as they 'run together'; thus the *n* in *input* will sound like an *m*. Stresses will be added: the word *the* will be unstressed; *input* and *data* will be stressed on their first syllables. Phonological rules describe how actual sounds (phones) are derived from phonemes.

The levels are to some extent THEORETICAL CONSTRUCTS, and what counts as belonging to a particular level depends partly on the theory used (as indeed does the terminology used). For example, the following sequence of words is clearly in error at the syntax level:

**Green sleep ideas colourless furiously*

(Note the convention that erroneous NL is preceded by *.) However, the sequence:

**Colourless green ideas sleep furiously*

can be considered to be in error at the syntax level or the semantic level or both, depending on the theory used. In this module, syntax is generally interpreted fairly narrowly, so that I would say that the second sequence is syntactically valid but semantically anomalous. On the other hand, some linguists, including the originator of this example, Chomsky, have said that it is syntactically anomalous.

Linguists naturally attach considerable importance to their theories, and debates as to the merits of alternative approaches can be fierce. As computer scientists, my view is that we should take an 'engineering' approach, asking only which theory, if any, is the most useful in designing programs to carry out natural language processing.

The following table may help in mastering some of the terms used in discussing these levels.

Levels		*Adjective:*	*Entities:*	*Adjective:*
Meaning:	Pragmatics	pragmatic		
	Semantics	semantic		
Grammar:	Syntax	syntactic		
	Morphology	morphological	morpheme	morphemic
Sounds:	Phonology	phonological	phoneme	phonemic
	Phonetics	phonetic	phone	phonetic

I use the term **Natural Language Processing** (NLP) to cover all levels of processing. However, it is often used in a way which excludes speech, covering only processing which starts from the written word. It is then necessary to use the term 'Speech and Natural Language Processing' (SNLP) to cover all levels. The inter-disciplinary nature of NLP means that terminology in general is rather variable; you need to remember this when using material by different authors.

4 Dialects and Language Change

One of the striking features of natural languages is the way in which they change. Changes in word usage tend to be noticed in an individual's lifetime. Thus I am aware of the change in the meaning of the word *gay*, and can give a clear explanation of what has been involved. Syntax changes are more obvious in hindsight. If I read a book written before the Second World War, I may be conscious of an 'old-fashioned' feel to the sentence construction, but will probably find it difficult to specify exactly what causes this. Reading Shakespeare on the other hand reveals obvious syntax differences. Any truly 'intelligent' natural language processing system will have to cope with at least the pace of lifetime language changes.

In many cases, it seems that historical changes in languages were caused by the spread of what were originally dialect variations. British speakers of English are often aware of the way in which American usages can spread, eventually displacing the original English usage. One which seems to have occurred in my lifetime is the use of *billion*, now almost always used with the American meaning of a thousand million but when I was a child used in Britain to mean a million million. Thus it may be that an NLP system which can cope with dialect differences will automatically cope with language changes.

Dialect differences and language changes often arouse strong emotions. In studying language for the purposes of natural language processing, our object is to DESCRIBE not PRESCRIBE. For example, in order to specify a particular set of syntax rules or a particular pronunciation, I will sometimes need to use terms like 'Standard English English' (SEE)[2] or 'Standard American English' (SAE). This does not in any way imply that SEE is linguistically better (or even more standard) than any other dialect of English.[3]

---

[2] SEE is a very clumsy term, but I personally find the alternatives (such as 'Received Pronunciation' or 'BBC English') worse. 'British English' is widely used, but there at least as many differences between say 'Scottish English' and 'Northern Irish English' as between 'English English' and 'American English'.

[3] Indeed in many ways it is clearly neither. Consider the series of possessive pronouns: *my*, *your*, *his*, *her*, *its*, *our*, *your*, *their*. From these we might logically expect the reflexive pronouns: *myself*, *yourself*, *hisself*, *herself*, *itself*, *ourselves*, *yourselves* and *theirselves*. In many non-prestigious dialects of English, this is

5 Books

Except for the continuous assessment (see the separate handout), the lecture notes are intended to be a sufficient (albeit condensed) source of information. If you need additional linguistics background material I recommend:

Atkinson, M., Kilby, D.A. & Roca, I. (1988) *Foundations of General Linguistics* (2nd ed.), London: Unwin Hyman. [P 121]

A livelier alternative (but American and so NOT to be used for phonetics!) is:

Fromkin, V. & Rodman, R. (1993) *An Introduction to Language* (5th ed.), Holt, Rinehart & Winston. [P 106]

Probably the most comprehensive current textbook for NLP is:

Jurafsky, D. & Martin, J.H. (2000) *Speech and Language Processing*, Prentice Hall.

Again it has the problem that the speech sections are oriented to American English. It's an excellent book to use to supplement my handouts, especially if you want to find out more about those areas of NLP that I'm NOT covering in the module.

Some older 'standards' on NLP in an AI context are:

Gazdar, G. & Mellish, C. (1989) *Natural Language Processing in Prolog: an introduction to computational linguistics*, Addison-Wesley [QA 76.7.P76]
Covington, M.A. (1994), *Natural Language Processing for Prolog Programmers*, Prentice-Hall [QA 76.7.P76]
Allen, J. (1995) *Natural Language Understanding* (2nd ed.), Benjamin/Cummings. [P 121]

Personally I find them less useful than might be expected. Covington and Gazdar & Mellish are, as their titles suggest, oriented towards practical Prolog programming. Allen is perhaps more comprehensive, but is strongly oriented towards Lisp and towards predicate calculus for the semantic representation. (It is also very parochial in its total neglect of languages other than English.)

A comprehensive overview of natural language from a more psycholinguistic viewpoint is given by:

Jackendoff, R. (2002) *Foundations of Language*, Oxford UP.

Jackendoff's "post-Chomskyan" perspective isn't universally shared, but the tradition he represents has been very influential in NLP.

For reference only, some other books which have influenced me in preparing these lectures include:

Aitchison, J. (1978) *Linguistics* (2nd ed.), Sevenoaks: Teach Yourself Books. [P 121]
Cook, V. J. (1988) *Chomsky's Universal Grammar: An Introduction*, Oxford: Blackwell. [P 85.C5]
Huddleston, R.D. (1976) *An Introduction to English Transformational Syntax*, Longman. [PE 1361]
Radford, A. (1997) *Syntactic Theory and the Structure of English*, Cambridge University Press. [P 291]
Lyons, J. (ed.) (1970) *New Horizons in Linguistics*, Penguin. [P 121]
Rich, E. & Knight, K. (1991) *Artificial Intelligence* (2nd ed.), McGraw-Hill. [Q 335]
Strang, B. M. H. (1970) *A History of English*, Methuen (reprinted by Routledge, 1989). [PE 51]

The data in §2 is taken from:

Comrie, B., Matthews, S. & Polinsky, M. (eds) (1977) *The Atlas of Languages*, Bloomsbury.
Crystal, D. (1987) *The Cambridge Encyclopedia of Language*, CUP.

---

exactly the series that is found. But in the 'standard' dialect we must use instead *himself* and *themselves* – but on no account *meself*!