Language: survey

vishvAs vAsuki

February 28, 2012

Contents

Part I

General structure

1 Themes

Computational linguistics tries to understand human language generation and comprehension using computer models.

1.1 Human vs programming languages

Natural language processing considers how human language can be generated or understood by a computer.
Design of language which can be translated to instructions understood by a compiler is considered elsewhere.

1.2 Descriptive vs prescriptive

Descriptive linguistics studies the language as it is actually used by a certain group of people. Prescriptive linguistics considers the language as it 'ought' to be used.

1.3 Example applications/ problems

Authorship attribution.
Interaction with humans: Eg: IBM built 'Watson', a program which participates in the propular TV trivia show jeopardy, where it responds to phrases with questions to which they are an answer.
Human language acquisition.

1.4 Ambiguity: main challenge

Comprehension and generation of text is just translation between natural language and the language of logic. The ambiguity in natural language is the main problem in this process.

1.5 Following research

ACL¿ NACL (Held only when ACL is outside north america)¿ COLING.

2 Language structure

2.1 Sounds words sentences

There are sounds. Certain ordered sounds make up words. Words make up phrases (Eg: noun phrases, verb phrases), which in-turn make up sentences.

2.2 Word morphology

2.2.1 Morpheme

Parts of a word which are not further divisible are called morphemes. Eg: In 'sunism', 'sun' is the root/ stem word, while 'ism' is an affix; both are morphemes.

2.2.2 Inflection

Inflectional morphemes alter the words gender, tense, number etc.: 's' in suns. 'ism' (as in 'sunism') is a derivational morpheme as it is used to derive a new separate word.

2.3 Word/ phrase roles/ categories

The general role a word/ phrase plays (as against its specific meaning) in a sentence is called a 'part of speech'.
Depending on the number of words that can play a certain role, the word role/ class is closed or open. Eg: determiners in English is a closed class.

2.3.1 Content vs function words

Words are either content words which have lexical meaning, or are function words (aka grammatical / structure-class words) which serve grammatical purpose while carrying little or no independent meaning.

2.3.2 Parts of speech

- noun (communicates objects/ subjects)
- verb (communicates action), article (a, the ..)
- adjective (qualifies the noun)
- adverb
- preposition and postposition (Aka adposition, clarifies a noun's grammatical case: of, on, in etc..)
- pronoun (a generic noun, like 'he' ..)
- conjunction (and, but, or)
  A subordinating conjunction, which joins a subordinate clause to the main clause (Eg: if).
- particle: A function word which does not belong to other classes.
  interjection and filler words('oh!').

Other important parts of speech include wh-word (question word), Modal (could, would, must, can, might ..), list item markers.

2.3.3 Noun and verb sub-roles

A noun word can have further generic sub-roles based on what gender, number and grammatical case (especially possessive in English) it communicates.
A verb word can also have sub-roles based on whether it communicates grammatical/ subject's gender, subject's number, person (1st, 2nd or 3rd person), aspect, mood (imperative, wishfulness), time and completeness of the action (present/ past/ future) etc.. This is called 'tense'. If it calls attention to the completeness of an action it is called 'perfect'.

2.4 Semantics and ambiguity

2.4.1 Word meaning ambiguity

Meanings of words themselves are ambiguous. Even their part-of-speech can be ambiguous - this ambiguity can lead to different parse trees; for example 'dogs' in: 'The sailor dogs the barmaid'.

2.4.2 Sentence ambiguity

Meanings of sentences, even after the meanings of words have been fixed, is still ambiguous due to the presence of differing possible parse trees. Then, there is analogy, sarcasm etc..

2.5 Grammar

Language structure is described by a grammar. A grammar attempts to model how a sentence of the language is formed.
A grammar can be descriptive or generative, the distinction being that in the latter case the decision made by the grammar of sentence correctness/ membership is ultimate. Generative grammars include computer languages and Sanskrit.

3 Language and thought

3.1 Complex thought and language

Complex thought about a certain topic requires suitable vocabulary; complex ideas become simple in science when the right theory/ notation/ framework is in place.

3.1.1 Examples/ evidence

The first generation of deaf-mutes in a central American country were unable to put themselves in others' place: they failed the test where they are asked to

guess where a person would look for a thing whose position has been changed in their absence. This is seem among children below a certain age.
A certain Hispanic deaf-mute man did not know that objects had names or symbols. So, to communicate or comprehend simple ideas: such as reliving the memory of a bull-fight, took a long time: 45 minutes of miming.

3.2 World-view and language

Languages differ essentially in what they must convey and not in what they may convey. Your language changes the way you think and feel about many things - by making them more or less important in your world-view.

3.2.1 Examples

In saMskRRita, you are forced to convey the sex of the person you interact with, in English, you are often forced to specify the time of a meal; and inanimate objects have gender: this changes the way you feel about them.
In some languages, you specify geographic direction 'N/S/E/W', whereas in others you specify front/ back etc..: memories (even gestures) in such languages are tagged with geographic direction: the sense of direction in speakers of such languages is extremely keen. They have perhaps trained their brains to maintain an accurate bird's eye view of their location.
Our brains are trained to exaggerate the distance between shades of color if these have different names in our language.
Some languages, like Matses in Peru, oblige their speakers, like the finickiest of lawyers, to specify exactly how they came to know about the facts they are reporting.

3.3 Social interactions

Vagueness and indirectness in language is used in communicating an idea while leaving a slight window of deniability. Eg: Idioms such as 'will you come up to see my etchings?'.

4 Models of communication

4.1 Speech generation

Speech generation involves the following tasks:

- Intention: Generating the thought to be spoken in the internal language of logic.
- Generation: Translating the logical language into natural language.
- Synthesis: Speaking the generated words with appropriate stresses, accents etc..

4.2 Speech comprehension

4.2.1 Tasks

Speech comprehension involves the following tasks:

- Perception: Translating the sounds heard or symbols seen to words or tokens; this is akin to lexical analysis by compilers.
- Analysis: Inferring a logical sentence equivalent to the spoken words. This involves the following tasks:

  Parsing/ syntactic tasks: Understanding the phrasal structure, for example using parse trees. As part of this process, the following tasks may be done:

  Part of speech tagging may be done.

  Phrase chunking: In this task, words are chunked to form entities such as a noun phrase and a verb phrase, by collecting together qualifiers (adjectives/ adverbs) and prepositional phrases.

  Semantic interpretation: Translating the words spoken to the language of logic. As part of this, the following tasks show up:

  Word sense disambiguation.

  Semantic role labeling: Here, a noun phrase's relation to the verb is decided. In case of Indic languages (esp Sanskrit), doing this is simplified due to the kAraka system.

  Pragmatic interpretation: Understanding what was meant as opposed to what was said. Eg: Detecting sarcasm.
- Incorporation: Relating the inferred logical statement to the knowledge base, and judging whether should be incorporated into the knowledge base and doing so if appropriate.

  A common task is resolving what entity a given common noun or pronoun refers to. Depending on whether the referred entity is within the corpus or whether it should be judged from deep understanding of the context, one classifies these into endophora and exophora.

There is ambiguity in every subtask above.

4.2.2 Comprehenders: Manual vs automatic building

4.2.2.1 Manual design

In this approach, a group of people use their knowledge of the language to build an expert system to comprehend a language. But, it is very hard for a group of people to encode all the rules of a language, for which there are a huge number of exceptions. Eg: 'is' in different tenses. An exception to this is probably pANini's grammar for saMskRRita.

Also, even people, in the face of ambiguity, make use of statistics/ probability in comprehension, eg: 'What do people commonly mean by this sentence?'

4.2.2.2 Automatic learning

Here, the computer learns the rules for comprehending a language by looking at statistics learned from annotated corpora.

4.2.3 Non-Modularity

Consider the various tasks involved in speech comprehension. These tasks may be viewed as happening in a sequence. In reality, however, they are intertwined: For example, one uses the syntactic structure to determine that the words in a sound correspond to 'I ate a rabbit', rather than 'Eye eight arab it.'.

4.3 Part of speech tagging

4.3.1 Problem

The task is to label all words in a sentence with the most appropriate part of speech. The set of appropriate parts of speech is usually fixed. It either taken to be the set defined by classical grammars of the language or produced more dynamically (supertagging?).

4.3.2 Purpose

Once part of speech tags are produced, they can be used as a feature while building parse-trees.
It is also useful in word sense disambiguation.

4.3.3 Approaches

This is an instance of the sequence label prediction problem. So, the approaches described for that general problem in the probabilistic models survey apply here as well.

4.3.3.1 Performance ideal

Given the same text, due to ambiguity in natural language, even taggings produced by human taggers don't agree with each other fully. For example, in english, the level of agreement is around 97%. So, this forms an ideal for tagging algorithms.

4.3.4 Using a tagged sample

4.3.4.1 Results

Consider the following for the Portuguese-boscque corpus from CoNLL 2006.
Simply labeling every word with the most frequent tag (noun) yields around 60% accuracy. Labeling each word with the tag most likely to correspond

to that word yields around 92% accuracy. Using HMM yields around 95% accuracy.
OpenNLP's implementation of MAXENT model with event threshold 5 and 100 iterations yields around 96.5% accuracy.

5 Language models

5.1 Definition

A language model models the probability of every possible string being a sentence spoken by a human.

5.2 Applications

Language models are useful in speech recognition, machine translation, handwriting recognition, machine translation, speech generation, (context sensitive) spelling correction etc..

5.3 Multi-set of words model

This very simple model ignores word order - a very important information. It models $Pr(w_{1:m}) = \prod Pr(w_i)$; and the parameters of this model - the word occurrence probabilities - are easily estimated; and it is known to follow the Zipf's law, which is a heavy tailed distribution. This is actually the 1-gram model, a member of the n-gram model family described elsewhere.

5.4 n-gram model

See Probabilistic models survey.

5.4.1 Performance

A trigram model is very commonly used. Google, as of 2011, seems to have built a good 5-gram model.

5.5 Finite state transducer (FST)

5.5.1 Definition

These are Finite State Automata (FSA described in the boolean functions survey) extended to include production of outputs corresponding to each state transition.
It is specified by a state set S, a subset of initial and final states, a state transition relation: $S \times I \rightarrow S \times O$ - where I, O are the input and output alphabet.

Because of the non-deterministic nature of the transitions, multiple output strings may be produced for a single input string. A probabilistic version of an FST produces probability scores along with the output strings.
Depending on whether a non-empty output string is produced by the operation of the FST on a given input string s, it may be declared to be accepted or rejected.

5.5.2 Transition graph

As in the case of a FSA, FST can be written as a transition graph - except that each edge is labelled by both an input and an output character.
[**Incomplete**]

Part II

Languages

6 Activity language

A person's motion, as he goes about various tasks, can be modeled as a language. There is a hierarchy in motions, which is akin to the sound - word - sentence - hierarchy.

7 Japanese

7.1 Sounds

l is pronounced/ heard as r. T is heard as t.

7.2 Words

Pronouns: watashi anata.

7.2.1 Noun Inflections

watashi no: my. Tokyo kara: from Tokyo. Tokyo made: To tokyo. Tokyo ni: In Tokyo.

7.2.2 Verb

kire - kill. desu - is. There are no plural forms.

7.2.2.1 Adverbs

shite (hurry). gosaimas (very much). not - ni.

7.3 Sentences

7.3.1 Greetings/ politeness

konnichiwa (pre-dusk), oyasuminasai (night). Jamen (I go/ bye). sayonAra (good bye). arigato (thanks).

7.3.2 Questions and answers

Questioning appendage: des-ka?
Answering: hai. iye.

8 References

Ray Mooney's course notes.