Natural language processing: state of the art, prospects for the next few years

Alistair Knott

Types of structure found in language

- Speech / orthography
- Lexical semantics
- Syntax
- Compositional semantics
- Dialogue

Types of structure found in language

- Speech / orthography (physical symbols $\leftrightarrow$ words)
- Lexical semantics
- Syntax
- Compositional semantics
- Dialogue

Types of structure found in language

- Speech / orthography (physical symbols $\leftrightarrow$ words)
- Lexical semantics (words $\leftrightarrow$ word meanings)
- Syntax
- Compositional semantics
- Dialogue

Types of structure found in language

- Speech / orthography (physical symbols $\leftrightarrow$ words)
- Lexical semantics (words $\leftrightarrow$ word meanings)
- Syntax (word sequence $\leftrightarrow$ syntactic structure)
- Compositional semantics
- Dialogue

Types of structure found in language

- Speech / orthography (physical symbols $\leftrightarrow$ words)
- Lexical semantics (words $\leftrightarrow$ word meanings)
- Syntax (word sequence $\leftrightarrow$ syntactic structure)
- Compositional semantics (syn. structure $\leftrightarrow$ utterance meaning)
- Dialogue

Types of structure found in language

- Speech / orthography (physical symbols $\leftrightarrow$ words)
- Lexical semantics (words $\leftrightarrow$ word meanings)
- Syntax (word sequence $\leftrightarrow$ syntactic structure)
- Compositional semantics (syn. structure $\leftrightarrow$ utterance meaning)
- Dialogue (utterance $\rightarrow$ next utterance)

Speech / orthography

Speech / orthography

Almost a solved problem.

Lexical semantics

Lexical semantics

How to represent the meaning of the word *crinkly*?

Lexical semantics

How to represent the meaning of the word *crinkly*?

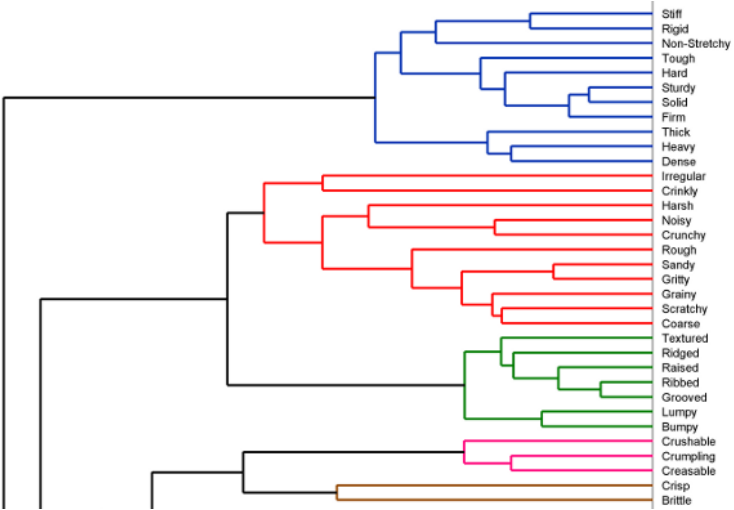
- Logic: $\lambda x[crinkly'(x)]$

Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used

Ilustrating the statistical modelling approach



Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used

Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used
- Embodied semantics: a tactile/auditory/sensory sensation

Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used
- Embodied semantics: a tactile/auditory/sensory sensation

How well can we map words to word meanings?

Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used
- Embodied semantics: a tactile/auditory/sensory sensation

How well can we map words to word meanings?

- The best methods use probabilities conditioned on nearby words.

Lexical semantics

How to represent the meaning of the word *crinkly*?

- Logic: $\lambda x[\mathit{crinkly}'(x)]$
- Stat. language modelling: the contexts in which *crinkly* is used
- Embodied semantics: a tactile/auditory/sensory sensation

How well can we map words to word meanings?

- The best methods use probabilities conditioned on nearby words.
  SEMEVAL/SENSEVAL compentitions:
  Coarse-grained word-sense disambiguation: over 90%.

Syntax

Syntax

People used to build grammars by hand.
Now, grammars are *learned*, from corpora of hand-parsed sentences.

Syntax

People used to build grammars by hand.
Now, grammars are *learned*, from corpora of hand-parsed sentences.

- To represent a realistic fragment of a natural language, a grammar must be *huge*.

Syntax

People used to build grammars by hand.
Now, grammars are *learned*, from corpora of hand-parsed sentences.

- To represent a realistic fragment of a natural language, a grammar must be *huge*.
- A large grammar finds that most sentences are massively ambiguous.
- To address this, grammars are probabilistic.
  The probability of each piece of structure is conditional on other structures nearby, and the words they contain.

Syntax

People used to build grammars by hand.
Now, grammars are *learned*, from corpora of hand-parsed sentences.

- To represent a realistic fragment of a natural language, a grammar must be *huge*.
- A large grammar finds that most sentences are massively ambiguous.
- To address this, grammars are probabilistic.
  The probability of each piece of structure is conditional on other structures nearby, and the words they contain.

PARSEVAL competition
Over 90% (metric involves precision, recall, tree-similarity)

What are syntactic structures?

What are syntactic structures?

Many of the structures in any given language are simply learned conventions.

What are syntactic structures?

Many of the structures in any given language are simply learned conventions.

- Some linguists think that syntactic structures are *nothing but* learned conventions.

What are syntactic structures?

Many of the structures in any given language are simply learned conventions.

- Some linguists think that syntactic structures are *nothing but* learned conventions.
- Others think that syntactic structures also reflect an *innate* capacity for language.

Compositional semantics

Compositional semantics

The syntactic structure of a sentence tells you *how to combine the meanings of its words*, to yield its overall meaning.

Compositional semantics

The syntactic structure of a sentence tells you *how to combine the meanings of its words*, to yield its overall meaning.

- We have quite good models of compositional semantics…

Compositional semantics

The syntactic structure of a sentence tells you *how to combine the meanings of its words*, to yield its overall meaning.

- We have quite good models of compositional semantics. . .
- Caveat 1: we don't know what syntax is.

Compositional semantics

The syntactic structure of a sentence tells you *how to combine the meanings of its words*, to yield its overall meaning.

- We have quite good models of compositional semantics…
- Caveat 1: we don't know what syntax is.
- Caveat 2: we don't know what word meanings are.

Aside: Sentence translation

Aside: Sentence translation

People used to build translation systems that found the meaning of the source sentence, and rendered it in the target language.

Aside: Sentence translation

People used to build translation systems that found the meaning of the source sentence, and rendered it in the target language.

Now, translation systems just use statistics computed over surface words.

Aside: Sentence translation

People used to build translation systems that found the meaning of the source sentence, and rendered it in the target language.

Now, translation systems just use statistics computed over surface words.

- They are trained on corpora of hand-translated sentences.

Aside: Sentence translation

People used to build translation systems that found the meaning of the source sentence, and rendered it in the target language.

Now, translation systems just use statistics computed over surface words.

- They are trained on corpora of hand-translated sentences.
- They compute conditional probabilities of words/phrases in L2 given words/phrases in L1.

Aside: Sentence translation

People used to build translation systems that found the meaning of the source sentence, and rendered it in the target language.

Now, translation systems just use statistics computed over surface words.

- They are trained on corpora of hand-translated sentences.
- They compute conditional probabilities of words/phrases in L2 given words/phrases in L1.
- Performance: see Google translate.

Dialogue

Dialogue

1. Question-answering systems. (A special case of dialogue systems)

Dialogue

1. Question-answering systems. (A special case of dialogue systems)

There are two basic methods:

- Text-based methods: rearrange the query sentence to create templates for answer sentences, then search the web for such sentences.
- Knowledge based methods: build a semantic representation of the query, then query a knowledge base.

Dialogue

1. Question-answering systems. (A special case of dialogue systems)

There are two basic methods:

- Text-based methods: rearrange the query sentence to create templates for answer sentences, then search the web for such sentences.
- Knowledge based methods: build a semantic representation of the query, then query a knowledge base.
- IBM's Watson uses a mixture of these methods.

Dialogue

1. Question-answering systems. (A special case of dialogue systems)

There are two basic methods:

- Text-based methods: rearrange the query sentence to create templates for answer sentences, then search the web for such sentences.
- Knowledge based methods: build a semantic representation of the query, then query a knowledge base.
- IBM's Watson uses a mixture of these methods.

QA systems are evaluated in the TREC competition.

- Circa 2013: The best systems can successfully answer 70% of 'factoid' questions.
- '90% of nurses follow Watson's guidance' in a specialised healthcare application.

Dialogue

Dialogue

2. Human-computer dialogue systems.

Dialogue

2. Human-computer dialogue systems.

Two basic methods:

Dialogue

2. Human-computer dialogue systems.

Two basic methods:

- Surface-based methods (e.g. Microsoft's Tay): basically like text-based machine translation.

Dialogue

2. Human-computer dialogue systems.

Two basic methods:

- Surface-based methods (e.g. Microsoft's Tay): basically like text-based machine translation.
- Knowledge-based methods: the best ones use dialogue managers, and often planning systems.
  These systems are clever, but fragile.

The future of NLP

The future of NLP

Some parts of the pipeline already work very well.

- Speech, writing
- Word-sense disambiguation
- Those aspects of syntax that reflect conventions
- Compositional semantics

The future of NLP

Some parts of the pipeline already work very well.

- Speech, writing
- Word-sense disambiguation
- Those aspects of syntax that reflect conventions
- Compositional semantics

Other parts still need big theoretical advances.

- Grounding lexical semantics in the world
- Answering foundational questions about ‘what syntax is’
- Building dialogue systems that are robust *and* deep.


I WISH I WAS
DEEP INSTEAD
OF JUST MACHO!