24.10.2014

Natural Language Processing

Charlie London

Description

What is it?

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (**natural**) languages. As such, NLP is related to the area of human–computer interaction.

NLP is very important in many automated services, such as automated phone lines (those you talk to), online chatbots like Cleverbot and automated assistants such as Siri and Google Now.

Uses

What can it do?

- Automatic Summarisation - used most often in applications that summarise news articles for quick and easy reading. NLP is necessary to make sure that the article still makes sense

- Machine Translation - a very difficult problem involving automatically translating text from one language to another. It is a problem termed ‘AI-complete’ meaning that it needs a large breadth of human knowledge and real world experience to work (e.g. grammar, semantics, colloquialism). This is why services such as Google Translate are often very unreliable

- Natural Language Generation - convert information from computer databases into language a a human can understand

- Optical Character Recognition - recognise and determine printed text inside an image e.g. computers could eventually solve CAPTCHAs

Uses

What can it do?

- Question Answering - make a computer able to answer a question given to it in human-language. Again very important for automated assistants on smartphones.
- Sentiment Analysis - used to identify opinions, especially on social media, usually about certain products, in order to be used by advertisers.
- Speech Recognition - determine the textual representation of a sound clip of someone speaking. Another AI-complete problem as there aren't many pauses between words natural speech, meaning the computer can create amalgamations of words. Speech segmentation also has to be used to prevent this from happening

The Basics

NLP and Machine Learning (covered sometime in the future)

- General learning algorithms learn grammatical rules through analysis of a large body of text called a corpora, which contains many typical real-world examples
- The corpora are sets of documents that have been hand-annotated (i.e. by a human) with the correct values for the machine to learn
- Modern NLP algorithms use a statistical approach which makes probabilistic decisions on what the input is, or is asking, by attaching weights or priorities of a certain value to each feature of the input section
- They can express how certain they are for a number of different possible answers and so give the most probable

The Basics

The Advantage of Machine Learning over Hand Written Rules

- Old NLP algorithms used hand-written rules, which were less accurate and more time consuming to create
    - Learning algorithms are less susceptible to tripping up when coming across unfamiliar inputs (words it hasn't seen before) and misspelled words. Hand-written rules only allow for certain things to happen, so you don't have the number of probable outcomes and it is more difficult to deal with these discrepancies
    - Systems that learn rules through analysing data are made more accurate simply by adding more data, which is a far more simple process than for hand-written systems. Hand-written systems are made more accurate by increasing the complexity of the rules, making the system more and more complicate and harder to manage effectively.

The Basics

Structures Used in Natural Language Processing

- Corpus - the main body of data that the algorithms learns from by analysing real world examples
  - Text corpus - large and structured set of texts, with which the computer can do statistical analysis and hypothesis checking to find the most probable answer
  - Speech corpus - database of speech audio files and text transcription. Can create acoustic models for use in speech recognition engines, for analysing speech inputs and converting them to text.

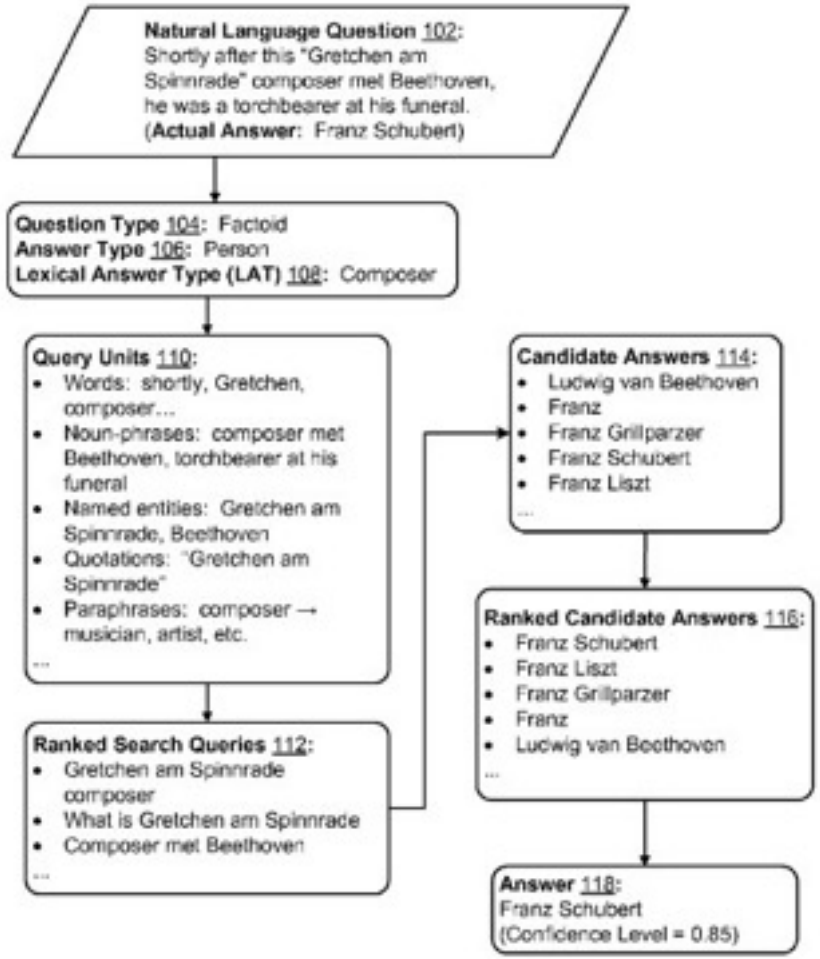
- Ontology - formal representation of a set of concepts within a domain and the relationship between those concepts, e.g. grammatical models

Process (Questions)

Question Understanding

- Analysis of natural language question to predict a question type and answer type:
  - Question types:
    - Factoid (e.g What's the capital of Mexico?)
    - Definition (e.g. What does antidisestablishmentarianism mean?)
    - Mathematical (e.g What is 4 factorial?)
    - Puzzle (e.g. What words can I make with the letters GAMRNAA?)
    - And more
  - Answer types include:
    - Person
    - Location
    - Quantity
    - Object
    - And more

Question Process Flowchart



Process (Questions)

Question Understanding

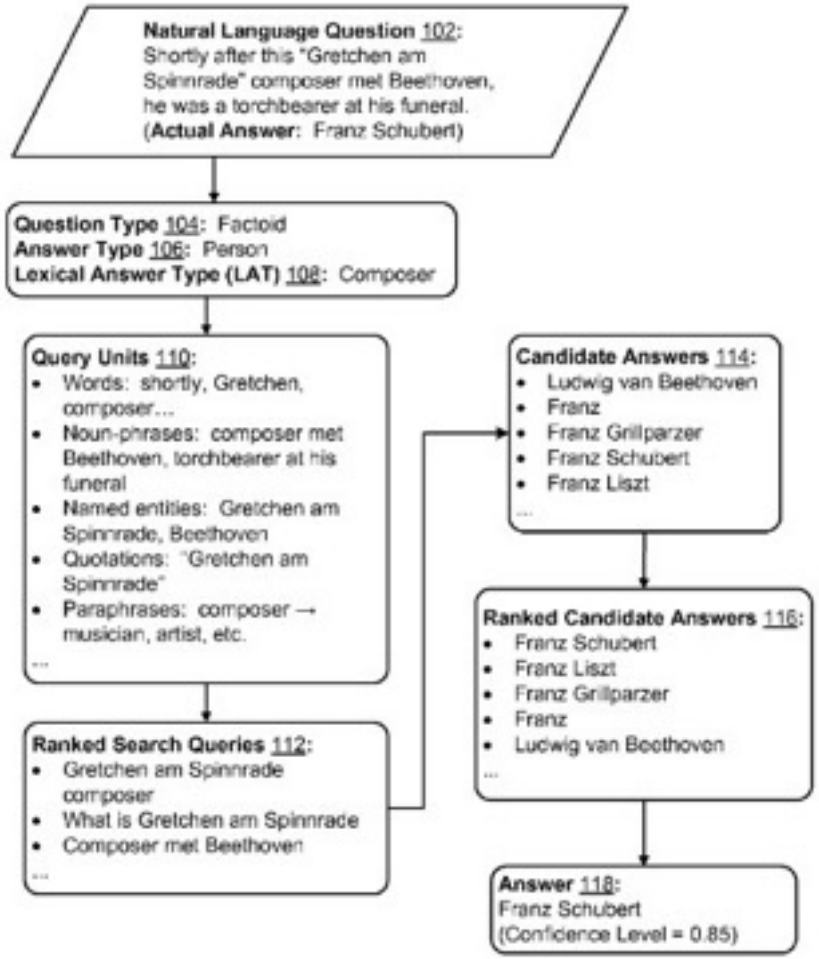
- Extracting query units from a natural language question e.g.:
  - Words, base-noun phrases, sentences, named entities, quotations, paraphrases, facts
  - Extraction Techniques:
    - Chunking - identifies parts of speech and short phrases (e.g noun phrases)
    - Sentence Boundary Detection - finds the end of sentences (as sometimes full stops signal other things like abbreviations)
    - Sentence Pattern Detection - detects what kind of sentence it is (e.g. subject + verb, subject+verb+object, etc.)
    - Parsing - analyses the string of symbols (natural language input) according to rules of formal grammar
    - Named Entity Recognition - locates and classifies elements in text into categories (e.g. person, place, mathematical value, etc.)
    - Part of Speech Tagging - reads text in natural language and assigns parts of speech to each word (e.g. adjective, verb, noun, etc.)
    - Tokenisation - breaks up text into words, phrases and symbols and other meaningful elements called tokens

Process (Questions)

Query Formulation

- Information from the question understanding phase is used to create search queries gather evidence and determine an answer, either from the internet or a separate database
- Uses query generation templates to create queries based on the question type, answer type, tokens, named entities, etc
- These queries can then be ranked and the top queries could be sent to the database or to the search engine being used, e.g Google for Google Now, Bing for Cortana
- Now that the question is better understood and queries have been sent, the next phase is entered

Question Process Flowchart

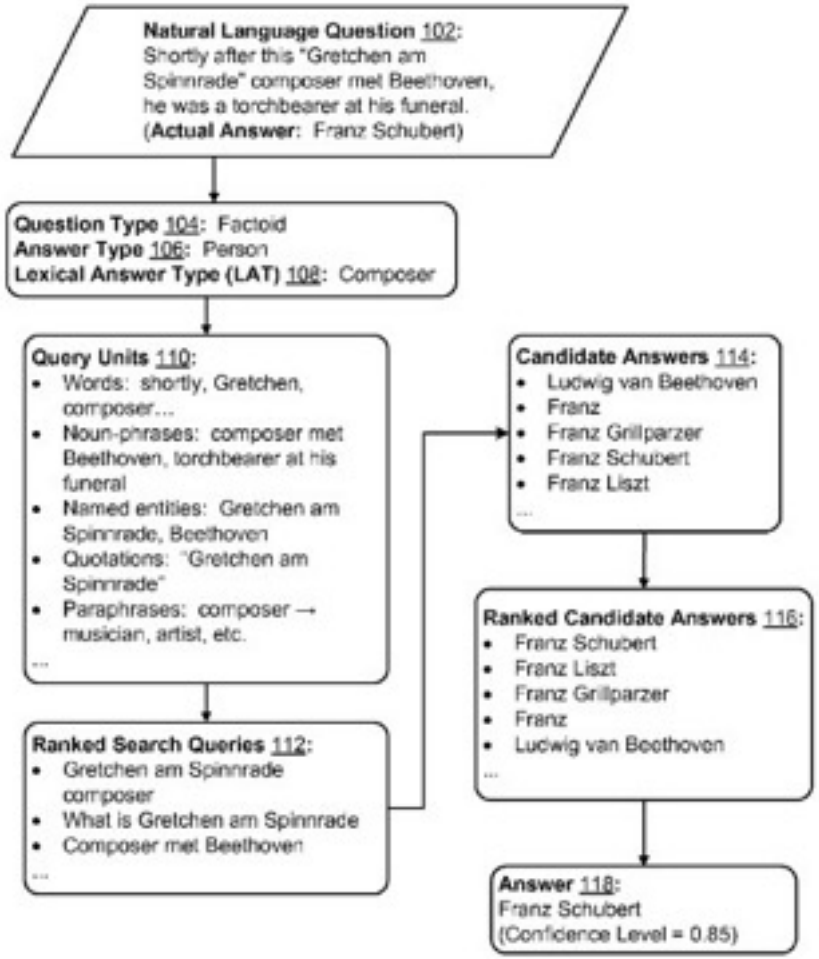


Process (Questions)

Evidence Gathering

- The top search queries are executed and the results are analysed
- The top results of each query are merged with one another to create a merged list of results
- Filtered to remove duplicate and nonsensical results

Question Process Flowchart



Process (Questions)

Answer Extraction/Ranking

- Candidate answers are extracted from the search results
  - Can involve dictionary-based entity recognition of entities that have a type that matches the answer type expected
- Candidate answers are then ranked by applying a set of features determined for the answer by a ranker that has been trained using a machine learning technique
- A confidence level can be determined for those top ranked answers
- That answer the computer is most confident with is given

Expansion Through Speech Recognition

- The functionality of this question answering could be increased by adding a speech to text input to allow the asking of questions by voice as well as text input

Speech Tools in C#

Programming and Speech Recognition

- In C#, simple speech recognition programs can be created using the System.Speech library and the SpeechRecognitionEngine
- Grammars have to be loaded into the SpeechRecognitionEngine or it won't recognise phrases
- A program can be written to tell the computer what to do when speech is recognised

```
1.  System.Speech
2.
3.  using System.Speech.Recognition;
4.  using System.Speech.Synthesis;
5.  using System.Threading;
6.
7.  SpeechRecognitionEngine
8.
9.  SpeechRecognitionEngine _recognizer = new SpeechRecognitionEngine();
10.
11. Grammar
12.
13. _recognizer.LoadGrammar(new Grammar(new GrammarBuilder("test")) { Name = "testGrammar" }); //
14. load a "test" grammar
```

And, just for some perspective

Processing Power

- The processing power of automated voice assistants such as Google Now and Siri is vast, and it comes with huge operating costs as well
  - It is estimated that Siri has three server farms in the USA alone, and at least one more for every other country it operates in.
  - “Apple’s lead cloud architect says that every instance of Siri runs on 32 powerful HP servers with a total of 1024 cores and 32 terabytes of RAM apiece. Specifically, each instance of Siri is made up of 4 HP c7k enclosures made up of 8 HP server blades each, with memory upgrades to 1TB of RAM.”
  - If one server dies it’s immediately replaced by another new one
  - It is estimated that the hardware in each server farm costs $1,000,000, and there will be other costs of maintenance, etc
  - However, this does;t seem as much when Apple made profits of $13 billion last year