Natural Language Processing 1

Katia Shutova

ILLC
University of Amsterdam

28 October 2020

Lecture 1: Introduction

Taught by...

Katia Shutova
Lecturer
e.shutova@uva.nl

Mario Giulianelli
Lab coordinator
m.giulianelli@uva.nl

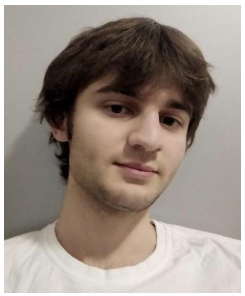
Christos Athanasiadis
Senior TA
c.athanasiadis@uva.n

Teaching assistants


Jaap Jumelet


Hannah Lim


Ece Takmaz


Oliviero Nardi


Massimo Spaconi


Christoph Hoenes


Omar Elbaghdadi


Anna Langedijk


Tamara Czinczoll

Overview of the course

- Introduction and broad overview of NLP
- Different levels of language analysis (word, sentence, larger text fragments)
- A range of NLP tasks and applications
- Both fundamental and most recent methods:
  - rule-based
  - statistical
  - deep learning
- Other NLP courses go into much greater depth

Course format

- The course will be offered (mostly) online:
  - **Recorded lectures** (released every week)
  - **Live Q & A** and **Discussion** on Wednesdays
  - **Exercises** and **practicals** (see below)
  - **Lab sessions**: mostly online.

- Limited on-site lab sessions:
  - three sessions per TA group
  - see Datanose for details
  - **do not come** to on-site sessions if you have any **COVID symptoms**!
  - masks available

Assessment

1. Practical assignments (**50%**)
   - Work in groups of 2
   - Implement several language processing methods
   - Evaluate in the context of a real-world NLP application — sentiment classification
   - Assessed by two reports (20% + 30%)
     - Practical 1: Mid-term report, deadline **13 November**
     - Practical 2: Final report, deadline **11 December**
2. Pen-and-paper exercises (individual work) (**20%**)
   - throughout the course
   - feedback from TAs
3. Exam on 18 December (**30%**)

Also note:

Course materials and more info:
`https://cl-illc.github.io/nlp1-2020/`

Contact

- Main contact – your TA (email on the website)
- Practicals – Mario: `m.giulianelli@uva.nl`
- Organisational – Chris: `c.athanasiadis@uva.nl`

Subject line should have **NLP1-20**

Sign up to groups by Fri, 30 October via the spreadsheet (link on Canvas)

Course Materials

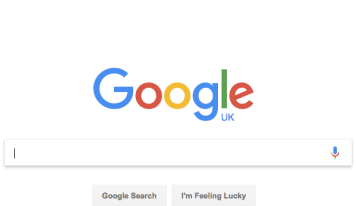
- Video lectures, slides, further reading, assignments posted on the **website**
- but... assignment submission will be via **Canvas**.
- **Piazza** for questions and discussion: `piazza.com/university_of_amsterdam/fall2020/nlp1`
- **Book:** Jurafsky & Martin, *Speech and Language Processing (2nd edition)*

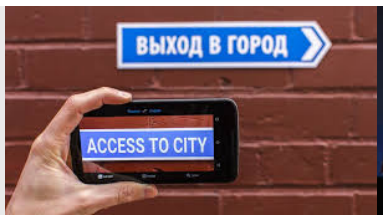
  3 edition (unofficial) at `https://web.stanford.edu/~jurafsky/slp3/`

What is NLP?

NLP: the computational modelling of human language.

Many popular applications



...and the emerging ones

Machine Translation

- Translate from one language into another
- Earliest attempted NLP application
- Early systems based on transfer rules, then statistical and now neural MT
- High quality with typologically close languages: e.g. Swedish-Danish.
- More challenging with typologically distant languages and low-resource languages

Retrieving information

- **Information retrieval**: return documents in response to a user query (Internet Search is a special case)
- **Information extraction**: discover specific information from a set of documents (e.g. companies and their founders)
- **Question answering**: answer a specific user question by returning a section of a document:

  What is the capital of France?
  Paris has been the French capital for many centuries.

Opinion mining and sentiment analysis

- Finding out what people think about politicians, products, companies etc.
- Typically done on web documents and social media
- More about this later today

Recently emerged applications

Automated fact checking

- classify statements and news articles as factual or not
- in an effort to combat misinformation

Abusive language detection

- automated detection and moderation of online abuse
- hate speech, racism, sexism, personal attacks, cyberbullying etc.

Other areas in which NLP is relevant

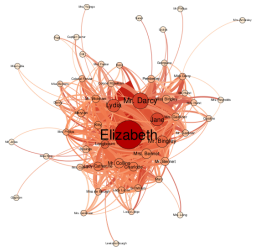
NLP and computer vision

- Caption generation for images and videos

The dog chewed at the shoes

Digital humanities

- e.g. social network in *Pride and Prejudice*



Computational social science

- analyse human behaviour based on language use (deeper than sentiment)

NLP and linguistics

1. Morphology — the structure of words: week 2.
2. Syntax — the way words are used to form phrases: weeks 1 and 2.
3. Semantics
   - Lexical semantics — the meaning of individual words: week 3.
   - Compositional semantics — the construction of meaning of longer phrases and sentences (based on syntax): week 4.
4. Pragmatics — meaning in context: week 5.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Why is NLP hard?

Synonymy and variability: different strings can mean the same or similar things

Did Google buy YouTube?

1. Google purchased YouTube
2. Google's acquisition of YouTube
3. Google acquired every company
4. YouTube may be sold to Google
5. Google didn't take over YouTube

Example from "Combined Distributional and Logical Semantics", Lewis & Steedman, TACL 2013

Wouldn't it be better if . . . ?

The properties which make natural language difficult to process are essential to human communication:

- Flexible
- Learnable, but expressive and compact
- Emergent, evolving systems

Synonymy and ambiguity go along with these properties.

Natural language communication can be indefinitely precise:

- Ambiguity is mostly local (for humans)
- resolved by immediate context
- but requires world knowledge

Wouldn't it be better if . . . ?

The properties which make natural language difficult to process are essential to human communication:

- Flexible
- Learnable, but expressive and compact
- Emergent, evolving systems

Synonymy and ambiguity go along with these properties.

Natural language communication can be indefinitely precise:

- Ambiguity is mostly local (for humans)
- resolved by immediate context
- but requires world knowledge

World knowledge...

"Knowledge is knowing that a tomato is a fruit"

BUT

"Wisdom is knowing not to put it in a fruit salad"

- Impossible to hand-code at a large-scale
- *either* limited domain applications
- *or* learn approximations from the data

Opinion mining: what do they think about me?

- Task: scan documents (webpages, tweets etc) for positive and negative opinions on people, products etc.
- Find all references to entity in some document collection: list as positive, negative (possibly with strength) or neutral.
- Construct summary report plus examples (text snippets).
- Fine-grained classification:
  e.g., for phone, opinions about: overall design, display, camera.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic design looks deceptively like metal. It feels solid in the hand and the build quality is great — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic ***design*** *looks deceptively like metal. It feels solid in the hand and the build quality is great — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic ***design*** *looks deceptively like metal. It feels solid in the hand and the* ***build quality*** *is great — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic ***design*** *looks deceptively like metal. It feels solid in the hand and the* ***build quality*** *is great — there's minimal give or flex in the body. It* ***weighs*** *149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic ***design*** *looks deceptively like metal. It feels solid in the hand and the* ***build quality*** *is great — there's minimal give or flex in the body. It* ***weighs*** *149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD ***display****, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.*

Sentiment classification: the research task

- Full task: information retrieval, cleaning up text structure, named entity recognition, identification of relevant parts of text. Evaluation by humans.
- Research task: preclassified documents, topic known, opinion in text along with some straightforwardly extractable score.
- Pang et al. 2002: *Thumbs up? Sentiment Classification using Machine Learning Techniques*
- Movie review corpus: strongly positive or negative reviews from IMDb, 50:50 split, with rating score.

Sentiment analysis as a text classification problem

c1:does the first word belong to class a or class b? eg.positive or negative setiment?

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, ..., c_J\}$
- *Output*:
  - a predicted class $c \in C$

predict the class of a new word

IMDb: An American Werewolf in London (1981)

Rating: 9/10

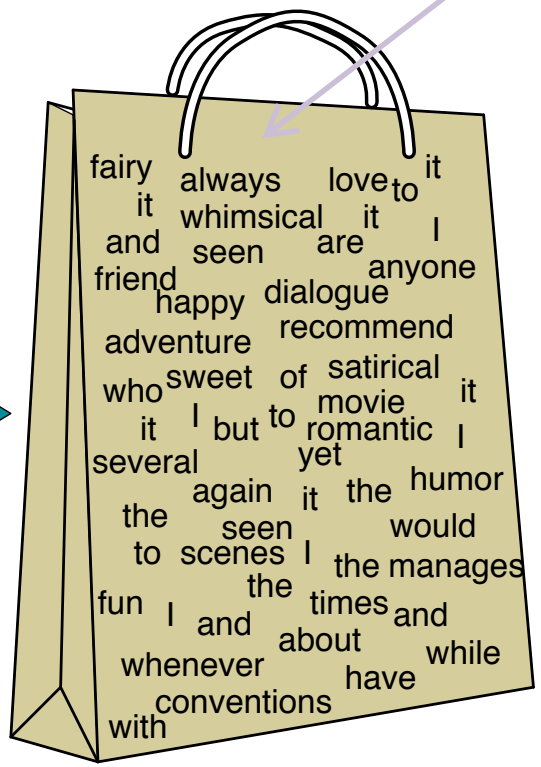
Ooooo. Scary.
The old adage of the simplest ideas being the best is once again demonstrated in this, one of the most entertaining films of the early 80's, and almost certainly Jon Landis' best work to date. The script is light and witty, the visuals are great and the atmosphere is top class. Plus there are some great freeze-frame moments to enjoy again and again. Not forgetting, of course, the great transformation scene which still impresses to this day.
In Summary: Top banana

Bag of words representation

the bag = a lexicon = a dictionary of the corpus

Treat the reviews as collections of individual words.

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Bag of words representation

- Classify reviews according to positive or negative words.
- Could use word lists prepared by humans — sentiment lexicons
- but machine learning based on a portion of the corpus (training set) is preferable.
- Use human rankings for training and evaluation.

Supervised classification

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, ..., c_J\}$
  - a training set of m hand-labeled documents $(d_1, c_1), ..., (d_m, c_m)$

document 1 is labeled as class c1

- *Output*:
  - a learned classifier $\gamma : d \rightarrow c$

map to

Classification methods

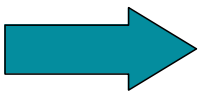
Many classification methods available

- Naive Bayes
- Logistic regression
- Decision trees
- k-nearest neighbors
- Support vector machines
- ...

Documents as feature vectors

The document d is represented as a feature vector $\vec{f}$:

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Naive Bayes classifier

input x
input:observation

Choose most probable class given a feature vector $\vec{f}$:

output: y

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|\vec{f})$$

this is a posterior
posterior
= p(原因|结果)
= p(model1|observation)
= p(class1|a feature vector)
= p(class1|a word)
= p(class1|a sentence)
= p(class1|a document)

observation可能由不同的model产生(model1, model2, model3...)，我们不知哪个model产生了该obs，所以所有的model，我们都要去计算概率。同理，we need to compute the prob for all class options, with given observation.

p(class1|observation)
p(class2|observation)
p(class2|observation)
goal: find the class i which gives the max prob

Apply Bayes Theorem:

$$P(c|\vec{f}) = \frac{P(\vec{f}|c)P(c)}{P(\vec{f})}$$

Constant denominator:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(\vec{f}|c)P(c)$$

Naive Bayes: feature independence

$$\hat{c} = \underset{c \in C}{\text{argmax}}\, P(\vec{f}|c)P(c)$$

Problem: need a very, very large corpus to estimate $P(\vec{f}|c)$

$$P(\vec{f}|c) = P(f_1, f_2, ..., f_n|c)$$

Conditional independence assumption ('naive'): assume the feature probabilities $P(f_i|c)$ are independent given the class c.

$$\hat{c} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1}^{n} P(f_i|c)$$

Naive Bayes: feature independence

likelihood prior

$$\hat{c} = \underset{c \in C}{\text{argmax}}\, P(\vec{f}|c)P(c)$$

Problem: need a very, very large corpus to estimate $P(\vec{f}|c)$

$$P(\vec{f}|c) = P(f_1, f_2, ..., f_n|c)$$

Conditional independence assumption ('naive'): assume the feature probabilities $P(f_i|c)$ are independent given the class c.

$$\hat{c} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1}^{n} P(f_i|c)$$

Naive Bayes: Learning the model

Maximum likelihood estimation: use frequencies in the data

$$\hat{P}(c) = \frac{Doccount(c)}{N_{doc}}$$

$$\hat{P}(f_i|c) = \frac{count(f_i, c)}{\sum_{f \in V} count(f, c)}$$

vocabulary:

the vocabulary is the set of words over all our data. so all the words in all negative and positive reviews.

does vocabulary means : all the words which appear in all the reviews? for example, if " fantastic" appear 3 times, then we record "fantastic" into vocabulary. vocabulary does not care how many times this word appears, it only record the word itself. right? Yes, correct!

f is a word in the feature vector

Problem with maximum likelihood

an element in the likelihood: p(input|class)
input can be a word or a sentence or a document or a review

What if we have seen no training documents with the word ***fantastic*** and classified as **positive**?

$$\hat{P}(fantastic|positive) = \frac{count(fantastic, positive)}{\sum_{f \in V} count(f, positive)} = 0$$

Zero probabilities cannot be conditioned away, no matter the other evidence!

prior

likelihood

$$\hat{c} = \underset{c \in C}{\text{argmax}} \, P(c) \prod_{i=1}^{n} P(f_i|c)$$

Laplace smoothing for Naive Bayes

Smoothing is a way to handle data sparsity

Laplace (also called "add 1") smoothing:

$$\hat{P}(f_i|c) = \frac{count(f_i, c) + 1}{\sum_{f \in V} (count(f, c) + 1)} = \frac{count(f_i, c) + 1}{\sum_{f \in V} count(f, c) + |V|}$$

a single element in the likelihood

Log space

max posterior
is converted to
max log of posterior

Use log space to prevent arithmetic underflow.

- Multiplying lots of probabilities can result in floating-point underflow
- sum logs of probabilities instead of multiplying probabilities

$$\log(xy) = \log(x) + \log(y)$$

- class with the highest log probability score is still the most probable

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}(\log P(c) + \sum_{i=1}^{n} \log P(f_i|c))$$

Test sets and cross-validation

validation set
https://www.youtube.com/watch?v=WqOtaCUCSIA&list=PLLssT5z_DsK8HbD2sPcUIDfQ7zmBarMYv&index=30

Divide the corpus into

- training set — to train the model
- development set — to optimize its parameters
- test set — kept unseen to avoid overfitting

or...

use **cross-validation** over multiple splits

- divide the corpus into e.g. 10 parts
- train on 9 parts, test on 1 part
- average results from all splits

Evaluation

Accuracy:

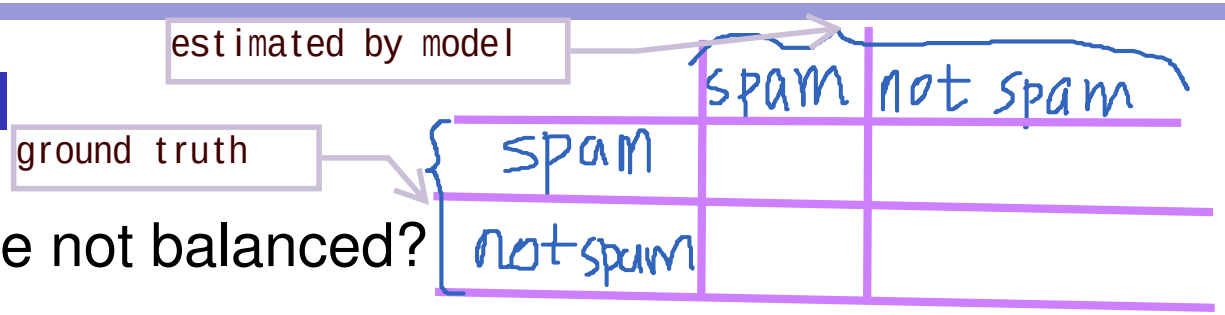
分子=sum of diagonal elements in the confusion matrix

$$Accuracy = \frac{Number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances}$$

Pang et al. (2002):

- The corpus is artificially balanced
- Chance success is 50%
- Bag-of-words achieves an accuracy of 80%.

Precision and Recall



What if the corpus were not balanced?

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

gold standard labels

system output labels		gold positive	gold negative	
	system positive	**true positive**	**false positive**	**precision** $= \frac{tp}{tp+fp}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{tp}{tp+fn}$		**accuracy** $= \frac{tp+tn}{tp+fp+tn+fn}$

F-measure

Also called *F-score*

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

β controls the importance of recall and precision

$\beta = 1$ is typically used:

$$F_1 = \frac{2PR}{P + R}$$

Error analysis

Bag-of-words gives 80% accuracy in sentiment analysis

Some sources of errors:

- Negation:

  Ridley Scott has never directed a bad film.

- Overfitting the training data:
  e.g., if training set includes a lot of films from before 2005, *Ridley* may be a strong positive indicator, but then we test on reviews for 'Kingdom of Heaven'?
- Comparisons and contrasts.

Contrasts in the discourse

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

More contrasts

AN AMERICAN WEREWOLF IN PARIS is a failed attempt . . . Julie Delpy is far too good for this movie. She imbues Serafine with spirit, spunk, and humanity. This isn't necessarily a good thing, since it prevents us from relaxing and enjoying AN AMERICAN WEREWOLF IN PARIS as a completely mindless, campy entertainment experience. Delpy's injection of class into an otherwise classless production raises the specter of what this film could have been with a better script and a better cast . . . She was radiant, charismatic, and effective . . .

Doing sentiment classification 'properly'?

- Morphology, syntax and compositional semantics
  - model relationships between words in the sentences, compose the meanings of phrases, negation, tense ...
- Lexical semantics
  - are words positive or negative in this context? Model word senses (e.g., *spirit*)
- Pragmatics and discourse structure
  - relationships between sentences; what is the topic of this section of text; co-reference resolution.
- Getting all this to work well on arbitrary text is very hard.
- Ultimately the problem is AI-complete, but can we do well enough for NLP to be useful?

Human translation?

Human translation?

I am not in the office at the moment. Please send any work to be translated.

Sentiment analysis practical: Part 1

Sentiment classification of movie reviews

1. Sentiment classification with a sentiment lexicon
2. Implement Naive Bayes classifier with bag-of-word features
3. Model grammar: word order and part of speech tags
4. Experiment with support vector machines (SVM) classifier
5. Evaluate and compare different methods

Assessed by the mid-term report, deadline 13 November

Sentiment analysis practical: Part 1

Sentiment classification of movie reviews

1. Sentiment classification with a sentiment lexicon
2. Implement Naive Bayes classifier with bag-of-word features
3. Model grammar: word order and part of speech tags
4. Experiment with support vector machines (SVM) classifier
5. Evaluate and compare different methods

Assessed by the mid-term report, deadline 13 November

Sentiment analysis practical: Part 2

- Experiment within a deep learning framework
- Include semantics
- Model the meaning of words, phrases and sentences
- Evaluate in the sentiment classification task

Assessed by the final report, deadline 11 December

Sentiment analysis practical: Part 2

- Experiment within a deep learning framework
- Include semantics
- Model the meaning of words, phrases and sentences
- Evaluate in the sentiment classification task

Assessed by the final report, deadline 11 December

Acknowledgement

Some slides were adapted from Ann Copestake, Dan Jurafsky and Tejaswini Deoskar