LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

15 November 2016

Class Outline

- NLP Applications: Information Extraction and Machine Translation
- Information Extraction
- Machine Translation
- Reminder: Assignment 5 submission due - tonight.

Regarding Assignment 5 - Question 1

- If you started working on A5 in the past 1-2 days, you may have noticed something is wrong with Stanford Parser online demo.
- Here are three options to answer Question 1 in this case:
  1. Use any other parser demo online and do the question 1
  2. use Stanford parser GUI that comes with the download (refer to README to know how)
  3. Write python code to print Stanford parser outputs and analyse that. (which you perhaps did in question 3)

NLP applications - overview

- *Text classification* - we discussed a few weeks back.
- *Information extraction*
- *Machine Translation*
- Information retrieval (Search)
- Dialog systems/conversational agents
- Question Answering/Summarization

Information Extraction

Information Extraction - overview

- Task: Extract different types of information (names, dates, relationships etc) from text.
- Let us take an example text:

```
Supermoon is an event that happens when a full moon is closest to Earth.
It orbits our planet in an oval shape so sometimes it comes closer to us
than at other times. To us Earth-lings, the moon appears 30 per cent brighter
and 14 per cent bigger.  By the way, supermoon is not an astrological term.
It's scientific name is perigee-syzygy, but supermoon is more catchy.
Astrologer Richard Nolle first came up with the term and he defined it as
"… a new or full moon which occurs with the moon at or near (within 90% of)
its closest approach to Earth in a given orbit", according to earthsky.org.

When is the next supermoon?
Monday, November 14. This supermoon will be the biggest and brightest in
70 years, so it will definitely be worth a look.  The "undeniably beautiful"
astronomical event will not come again until November 25, 2034, NASA said.
```

Information Extraction - tasks

- identify and classify "named entities" in the text - Named Entity Recognition (Supermoon, perigee-syzygy, Richard Nolle, NASA)
- Not sufficient to just say something is a proper noun. What sort of proper noun is it? person, event? organization? place?

Information Extraction - tasks

- identify and classify "named entities" in the text - Named Entity Recognition (Supermoon, perigee-syzygy, Richard Nolle, NASA)
- Not sufficient to just say something is a proper noun. What sort of proper noun is it? person, event? organization? place?
- Event detection. Key events in our example: supermoon, its next occurrence, etc.

Information Extraction - tasks

- identify and classify "named entities" in the text - Named Entity Recognition (Supermoon, perigee-syzygy, Richard Nolle, NASA)
- Not sufficient to just say something is a proper noun. What sort of proper noun is it? person, event? organization? place?
- Event detection. Key events in our example: supermoon, its next occurrence, etc.
- Temporal Expressions: Monday, November 14, November 25, 2034 etc.

Information Extraction - tasks

- identify and classify "named entities" in the text - Named Entity Recognition (Supermoon, perigee-syzygy, Richard Nolle, NASA)
- Not sufficient to just say something is a proper noun. What sort of proper noun is it? person, event? organization? place?
- Event detection. Key events in our example: supermoon, its next occurrence, etc.
- Temporal Expressions: Monday, November 14, November 25, 2034 etc.
- All this information is useful for: Relation extraction (identifying that Richard Nolle is the scientist who coined the term Supermoon)

Information Extraction - Methods

- Regular expressions (If you know the patterns of named entities, temporal expressions etc)
- Machine learning (If we do not know the patterns)
- Example: NER can be seen as a sequence labeling problem like in POS tagging, coupled with gazetteers containing names of persons, organizations etc.

Information Extraction and NLTK

- NER example (NER.py)
- Relation extraction example (RelExtraction.py)

Some current IE projects: code and data

- openIE - http://knowitall.github.io/openie/
- reVerb project - http://reverb.cs.washington.edu/
- Google relation extraction corpus https://research.googleblog.com/2013/04/50000-lessons-on-how-to-read-relation.html

Machine Translation

Machine Translation

- Task: translate text from one language to another (text can be a word, a sentence, a paragraph, a full document)
- Primary issues in solving the task:
    1. Different scripts, different word orders, different orthographic rules (no caps, no punctuation etc), different morphological structure

Machine Translation

- Task: translate text from one language to another (text can be a word, a sentence, a paragraph, a full document)
- Primary issues in solving the task:
  1. Different scripts, different word orders, different orthographic rules (no caps, no punctuation etc), different morphological structure
  2. Understanding genre specific nuances (translating news versus translating a novel)
  3. advanced issues: stylistic and cultural differences, idioms, metaphors etc.

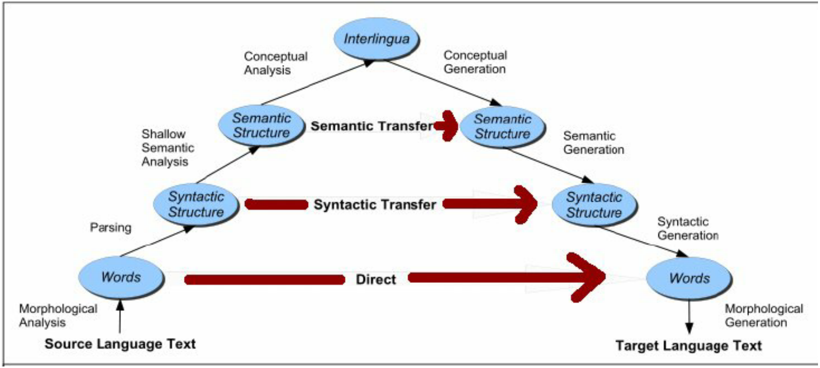
Machine Translation

- Task: translate text from one language to another (text can be a word, a sentence, a paragraph, a full document)
- Primary issues in solving the task:
  1. Different scripts, different word orders, different orthographic rules (no caps, no punctuation etc), different morphological structure
  2. Understanding genre specific nuances (translating news versus translating a novel)
  3. advanced issues: stylistic and cultural differences, idioms, metaphors etc.
  4. degrees of automatic translation: rough translation, translation + post editing, high-quality translation for very targeted domain specific language.

Machine Translation - Usage scenarios

- Rough translation: translating webpages for a general purpose reader
- Translation with post-editing: translation of software manuals for localization (Computer aided human translation). Computer does some part of translation followed by human translator edits.
- domain specific: weather forecasts (where vocabulary is limited, and language patterns are also limited).

Machine Translation - Classical view



source: https://goo.gl/slXqkS

Rule based MT

- Direct translation: do direct word by word translation from source to target. Not used now, but the same intuition of incremental translation works in currently used systems too.
- Transfer rules based translation: have rules for translating form X to Y, to account for word-order differences, and other such issues in addition to lexical transfer rules.
- Direct + transfer rules based translation

Machine Translation - Statistical view

Bayes is Back!

- Idea: If you have a large collection of parallel sentences from source and target languages, you can approximate human translation with a statistical model.
- If F is a foreign language and we are translating from F to English, probability of best translation:
  $= \text{argmax}_{E \in English}$ P(F|E)*P(E)
- Where P(E) is the language model for target language (English), which helps us choose the translation that is most likely in English language
- P(F|E) is the translation model. A commonly used translation model is "Phrase Based Statistical Machine Translation".
- Challenge: Creating alignments between the source and target sentences in the training data.

Machine Translation - Statistical view

An example phrase table. source: https://goo.gl/cORL3E

```
wife ||| manžele ||| 0.0107991 0.0203291 0.000693866 0.0012239 ||| 0-0 ||| 926 14412 10 ||| |||
wife ||| manželé ||| 0.0008 0.0032178 0.00062448 0.0025644 ||| 0-0 ||| 11250 14412 9 ||| |||
wife ||| jí ||| 0.000149016 0.0001522 0.000693866 0.0008742 ||| 0-0 ||| 67107 14412 10 ||| |||
wife ||| . manželka ||| 0.482759 0.584211 0.00194283 0.00513742 ||| 0-1 ||| 58 14412 28 ||| |||
wife ||| manželka , ||| 0.130435 0.584211 0.00187344 0.0347526 ||| 0-0 ||| 207 14412 27 ||| |||
wife ||| oženil ||| 0.0379747 0.0128655 0.00062448 0.0006411 ||| 0-0 ||| 237 14412 9 ||| |||
wife ||| manželského ||| 0.0205479 0.016385 0.0020816 0.001865 ||| 0-0 ||| 1460 14412 30 ||| |||
wife ||| stará ||| 0.00292208 0.0023866 0.00124896 0.0012239 ||| 0-0 ||| 6160 14412 18 ||| |||
wife ||| choť ||| 0.366492 0.370192 0.00485706 0.0044877 ||| 0-0 ||| 191 14412 70 ||| |||
wife ||| ženě , že ||| 0.031746 0.339074 0.00041632 0.000258566 ||| 0-0 ||| 189 14412 6 ||| |||
wife ||| manželství ||| 0.000212025 0.0001357 0.000485706 0.000408 ||| 0-0 ||| 33015 14412 7 ||| |||
wife ||| manžel ||| 0.000876534 0.0011182 0.000485706 0.0005828 ||| 0-0 ||| 7986 14412 7 ||| |||
wife ||| ženuška ||| 0.0740741 0.102804 0.00041632 0.0006411 ||| 0-0 ||| 81 14412 6 ||| |||
wife ||| ženatý ||| 0.00713558 0.0151643 0.000485706 0.0010491 ||| 0-0 ||| 981 14412 7 ||| |||
wife ||| žena mu ||| 0.186047 0.218609 0.000555093 0.000511047 ||| 0-0 ||| 43 14412 8 ||| |||
wife ||| wife ||| 0.75 0.6875 0.00041632 0.0006411 ||| 0-0 ||| 8 14412 6 ||| |||
wife ||| ženušku ||| 0.28 0.354839 0.000485706 0.0006411 ||| 0-0 ||| 25 14412 7 ||| |||
wife ||| se ženou ||| 0.0141988 0.31739 0.000485706 0.00281939 ||| 0-1 ||| 493 14412 7 ||| |||
wife ||| manžela ||| 0.00063012 0.0006776 0.000346933 0.0002914 ||| 0-0 ||| 7935 14412 5 ||| |||
wife ||| žena - ||| 0.0297619 0.218609 0.000346933 0.00279198 ||| 0-0 ||| 168 14412 5 ||| |||
wikileaks . ||| wikileaks . ||| 1 0.506273 1 0.529163 ||| 0-0 1-1 ||| 2 2 2 ||| |||
wikileaks founder julian ||| zakladatel wikileaks julian ||| 1 0.143037 0.5 0.118444 ||| 1-0 0-1 2-2 ||| 2 4 2 ||| |||
wikileaks founder ||| zakladatel wikileaks ||| 1 0.226288 0.5 0.20202 ||| 1-0 0-1 ||| 2 4 2 ||| |||
wikileaks has ||| wikileaks ||| 0.0555556 0.026732 0.5 0.277782 ||| 0-0 1-0 ||| 18 2 1 ||| |||
wikileaks ||| serverem wikileaks ||| 0.5 0.299373 0.047619 0.0308642 ||| 0-0 0-1 ||| 2 21 1 ||| |||
wikileaks ||| serveru wikileaks ||| 0.666667 0.298613 0.0952381 0.0385802 ||| 0-0 0-1 ||| 3 21 2 ||| |||
wikileaks ||| wikileaks , ||| 0.333333 0.597015 0.047619 0.175554 ||| 0-0 ||| 3 21 1 ||| |||
wikileaks ||| wikileaks se ||| 0.5 0.597015 0.047619 0.0234103 ||| 0-0 ||| 2 21 1 ||| |||
wikileaks ||| wikileaks ||| 0.722222 0.597015 0.619048 0.555556 ||| 0-0 ||| 18 21 13 ||| |||
wild and ||| přírodě a ||| 0.0454545 0.0861874 0.0135135 0.00991066 ||| 0-0 1-1 ||| 88 296 4 ||| |||
wild and ||| přírody a ||| 0.0115607 0.0270534 0.0135135 0.00786563 ||| 0-0 1-1 ||| 346 296 4 ||| |||
wild and ||| masa volně žijící a ||| 0.666667 0.128285 0.0135135 5.15859e-07 ||| 0-1 0-2 1-3 ||| 6 296 4 ||| |||
wild and ||| volně žijící a ||| 0.851852 0.128285 0.0777027 0.0100167 ||| 0-0 0-1 1-2 ||| 27 296 23 ||| |||
wild and ||| volně žijící i ||| 0.25 0.0259009 0.027027 0.000170571 ||| 0-0 0-1 1-2 ||| 32 296 8 ||| |||
wild and ||| divokým a ||| 0.5 0.249007 0.0168919 0.0136862 ||| 0-0 1-1 ||| 10 296 5 ||| |||
wild and ||| divokých a ||| 0.3125 0.216521 0.0337838 0.0278444 ||| 0-0 1-1 ||| 32 296 10 ||| |||
wild and ||| divoký a ||| 0.37037 0.309782 0.0675676 0.0346088 ||| 0-0 1-1 ||| 54 296 20 ||| |||
wild and ||| divoké a ||| 0.344262 0.300483 0.0709459 0.0504974 ||| 0-0 1-1 ||| 61 296 21 ||| |||
wild and ||| divoká a ||| 0.342857 0.313755 0.0405405 0.0292602 ||| 0-0 1-1 ||| 35 296 12 ||| |||
wild and ||| divoká , ||| 0.0714286 0.00498093 0.0135135 0.00101569 ||| 0-0 1-1 ||| 56 296 4 ||| |||
wild and ||| divocí a ||| 0.361111 0.353967 0.0439189 0.0110905 ||| 0-0 1-1 ||| 36 296 13 ||| |||
wild and ||| a ||| 9.243e-07 2.22931e-05 0.0168919 0.854051 ||| 1-0 ||| 5.4095e+06 296 5 ||| |||
wild and ||| divoce ||| 0.00230814 0.000508407 0.0135135 0.0083829 ||| 0-0 1-0 ||| 1733 296 4 ||| |||
wild and ||| divoce a ||| 0.0641026 0.0520312 0.0168919 0.0143154 ||| 0-0 1-1 ||| 78 296 5 ||| |||
wild and ||| přírodě , a ||| 0.117647 0.0861874 0.00675676 0.00313173 ||| 0-0 1-2 ||| 17 296 2 ||| |||
wild and ||| jsou divocí a ||| 0.4 0.353967 0.00675676 4.21995e-05 ||| 0-1 1-2 ||| 5 296 2 ||| |||
wild and ||| rozbouřené a ||| 0.285714 0.00852105 0.00675676 0.000157316 ||| 0-0 1-1 ||| 7 296 2 ||| |||
wild and ||| zběsilé a ||| 0.333333 0.0199711 0.00675676 0.000235974 ||| 0-0 1-1 ||| 6 296 2 ||| |||
wild and ||| divokou a ||| 0.153846 0.275975 0.00675676 0.0142368 ||| 0-0 1-1 ||| 13 296 2 ||| |||
wild and ||| zdivočelý a ||| 0.75 0.189787 0.0101351 0.00110121 ||| 0-0 1-1 ||| 4 296 3 ||| |||
wild and ||| mezi planými rostlinami a ||| 1 0.0382354 0.00675676 1.89327e-13 ||| 0-1 1-3 ||| 2 296 2 ||| |||
wild and ||| volně žijícími a ||| 1 0.220253 0.00675676 0.0013059 ||| 0-0 0-1 1-2 ||| 2 296 2 ||| |||
wild and ||| tak ||| 3.60929e-06 1.02675e-06 0.00337838 0.0034894 ||| 1-0 ||| 277063 296 1 ||| |||
wild and ||| divokého a ||| 0.25 0.305389 0.0101351 0.0174617 ||| 0-0 1-1 ||| 12 296 3 ||| |||
wild and ||| volně žijícího lososa a ||| 0.5 0.202648 0.0101351 2.57161e-09 ||| 0-0 0-1 1-3 ||| 6 296 3 ||| |||
wild and ||| planými rostlinami a ||| 1 0.0382354 0.00675676 2.83169e-10 ||| 0-0 1-2 ||| 2 296 2 ||| |||
wild and ||| volně žijících a ||| 0.666667 0.274843 0.00675676 0.0168332 ||| 0-0 0-1 1-2 ||| 3 296 2 ||| |||
wild and ||| zmatené a ||| 0.0666667 0.00510677 0.0101351 0.000314632 ||| 0-0 1-1 ||| 45 296 3 ||| |||
wild and ||| divoký kraj , i ||| 0.5 0.0625458 0.00675676 8.00791e-10 ||| 0-0 1-3 ||| 4 296 2 ||| |||
wild and ||| nespoutané a ||| 0.666667 0.0760806 0.00675676 0.00078658 ||| 0-0 1-1 ||| 3 296 2 ||| |||
wild and ||| volně žijící losos a ||| 1 0.128285 0.0101351 1.10183e-08 ||| 0-0 0-1 1-3 ||| 3 296 3 ||| |||
wild and ||| volně žijící exempláře a ||| 1 0.128285 0.00675676 2.404e-08 ||| 0-0 0-1 1-3 ||| 2 296 2 ||| |||
wild and ||| žijících a ||| 0.142857 0.37216 0.00337838 0.0922638 ||| 0-0 1-1 ||| 7 296 1 ||| |||
wild ||| volně žijícím ||| 0.966667 0.154382 0.00455474 0.00055449 ||| 0-0 0-1 ||| 30 6367 29 ||| |||
wild ||| divoké ||| 0.407031 0.403013 0.0818282 0.0591269 ||| 0-0 ||| 1280 6367 521 ||| |||
wild ||| volně žijícími ||| 0.876289 0.295407 0.0133501 0.00152906 ||| 0-0 0-1 ||| 97 6367 85 ||| |||
wild ||| volně žijícího ||| 0.664384 0.271795 0.0152348 0.00215077 ||| 0-0 0-1 ||| 146 6367 97 ||| |||
wild ||| divokými ||| 0.299492 0.336585 0.00926653 0.0063548 ||| 0-0 ||| 197 6367 59 ||| |||
wild ||| divokého ||| 0.368421 0.409594 0.0263861 0.0204458 ||| 0-0 ||| 456 6367 168 ||| |||
wild ||| divokém ||| 0.282486 0.330357 0.00785299 0.0068153 ||| 0-0 ||| 177 6367 50 ||| |||
wild ||| volně žijících ||| 0.439787 0.368625 0.0906235 0.0197098 ||| 0-0 0-1 ||| 1312 6367 577 ||| |||
wild ||| přírodě ||| 0.108025 0.115596 0.0164913 0.0116043 ||| 0-0 ||| 972 6367 105 ||| |||
wild ||| divokým ||| 0.277778 0.333973 0.0164913 0.0160251 ||| 0-0 ||| 378 6367 105 ||| |||
wild ||| divoká ||| 0.413686 0.420815 0.0417779 0.0342605 ||| 0-0 ||| 643 6367 266 ||| |||
wild ||| žijících ||| 0.322581 0.499149 0.0424062 0.108031 ||| 0-0 ||| 837 6367 270 ||| |||
wild ||| divokou ||| 0.360743 0.370143 0.0213601 0.0166697 ||| 0-0 ||| 377 6367 136 ||| |||
wild ||| divokej ||| 0.704545 0.729167 0.00486886 0.0032234 ||| 0-0 ||| 44 6367 31 ||| |||
wild ||| žijící ||| 0.0457574 0.106015 0.0161772 0.0642844 ||| 0-0 ||| 2251 6367 103 ||| |||
wild ||| wild ||| 0.800971 0.806867 0.0259149 0.0173144 ||| 0-0 ||| 206 6367 165 ||| |||
wild ||| přírody ||| 0.0371732 0.0362845 0.0142924 0.0092098 ||| 0-0 ||| 2448 6367 91 ||| |||
wild ||| divocí ||| 0.445736 0.474748 0.0180619 0.0129858 ||| 0-0 ||| 258 6367 115 ||| |||
wild ||| volně žijící ||| 0.704172 0.172058 0.0874823 0.0117284 ||| 0-0 0-1 ||| 791 6367 557 ||| |||
wild ||| divokých ||| 0.28122 0.290402 0.0463327 0.0326027 ||| 0-0 ||| 1049 6367 295 ||| |||
wild ||| divoce ||| 0.0525101 0.0697853 0.0142924 0.0167618 ||| 0-0 ||| 1733 6367 91 ||| |||
wild ||| divoký ||| 0.392814 0.415486 0.0515156 0.0405231 ||| 0-0 ||| 835 6367 328 ||| |||
wild ||| volně ||| 0.0550176 0.238101 0.039265 0.182446 ||| 0-0 ||| 4544 6367 250 ||| |||
wild ||| volné přírodě ||| 0.0584416 0.0602928 0.00141354 2.24439e-05 ||| 0-0 0-1 ||| 154 6367 9 ||| |||
```

Evaluating MT systems

- Human raters (in terms of: correctness, clarity, naturalness, grammaticality etc)
- automatic evaluation - BLEU score (ngram similarity based measure between a translated sentence and a gold standard human translation)
- Other automatic measures: TER, METEOR, NIST etc.

Useful Resources

- Chapter 25 in Jurafsky and Martin (very comprehensive overview)
- http://www.statmt.org/ - comprehensive website on readings, software, corpora related to developing and testing statistical machine translation systems.
- https://www.apertium.org - open source machine translation platform that lets you create rule-based MT models.
- Google Translate, Bing Translate etc - MT applications in real life.

Exercise in Analyzing machine translation

- Two popular machine translation software online: Google Translate and Bing translate
- Task: Try out how both these tools work for translating from English to your native language and viceversa (Native English speakers: choose another language you know. If you know only English, pair up with some other person).
- Spend some time with these and we can discuss your observations towards the end of the class.

Next class

- NLP for CALL - Introduction
- Assignment 5 discussion
- Final projects - status and discussion
- Recommended Readings: Burstein (2009), Chapelle and Chung (2010). Meurers (2013). Read atleast two of them. Large amount of thursday's class invovles discussion about these papers.
- Optional readings: Chapters on Writing Aids and Tutoring Systems in Language and Computers.