NLP Technologies Assignment 2

Deadline: May 1st

This document provides the exercises for the second module (Weeks 3, 4 and 5). We recommend you finish the Coreference component by the intermediate submission deadline (halfway Week 4).

We will ask you for each component to indicate how much time you spent on them. This is meant to ensure the workload of the course is appropriate (you get as much out of it as you can without becoming overworked).

PARTIAL VERSION: COVERING COMPONENT 1 and 2.1 ONLY. THERE WILL BE A COMPONENT 2.2 TO BE CARRIED OUT IN WEEK 5.

1 Coreference Resolution

In this component, we will work towards designing and building a coreference resolution system step-by-step. The overall learning goal is to acquire the skill of analyzing a task, data and possibilities offered by various technologies. The procedure should also give more insight into how to identify appropriate features for a machine learning setup. You can thus use this skill to improve your answers in Assignment 1.

The component is composed of the following steps:

1. Understanding the task and representation formats
2. Analyzing the phenomenon
3. Identifying relevant information
4. Designing part of a system (defining the features)

1.1 Understanding the task and representation

This component covers two important skills involved in NLP (and many other domains of artificial intelligence for that matter): (1) getting a first understanding of a task and (2) learning to deal with specific formats used to represent data.

We will use the following text (also provided to you in the coreference material package included in this module):

Coronavirus-infected Chinese tourist being treated in Thailand

A Chinese tourist was found to be infected with the new strain of coronavirus when she arrived in Thailand, is being treated in hospital and is expected to be discharged in a few days, Public Health Minister Anutin Charnvirakul said on Monday.

Mr Anutin said the 61-year-old woman was recovering at Bamrasnaradura Infectious Diseases Institute in Nonthaburi province.

She now had no fever or any respiratory symptoms. If doctors gave her a clearance she would be allowed to go home in a few days, said Mr Anutin.

Sixteen other people who were close to the woman on the same flight were examined, and the results were negative, he said.

Mr Anutin said 59 people in China have been confirmed infected with the new strain of the coronavirus, which has been linked to a sudden outbreak of pneumonia in central China. One of them died. All had attended big markets selling animals and seafood in Wuhan city. They were either workers or buyers. There had not been any human-to-human transmission of the virus.

The ill Chinese woman was the first person detected with the virus outside China. Her discovery and successful treatment was indicative of the efficiency and effectiveness of health services in Thailand, Mr Anutin said.

Health officials have been checking passengers from Wuhan arriving at Suvarnabhumi, Don Mueang, Phuket and Chiang Mai airports since Jan 3. They had found 12 ill passengers who justified being quarantined. Eight had so far been treated and discharged from hospital.

The Chinese woman was being was treated in an isolation ward. Her infection with the new coronavirus was confirmed on Sunday, Mr Anutin said.

The Public Health Ministry had not found anyone else infected with it, he said. One of Wuhan's largest meat and seafood markets was pinpointed as the centre of the mysterious pneumonia outbreak and was shut down on Jan. 1. The man who died had been a customer at that market. Chinese scientists identified the new virus strain last week.

1. Read the text carefully.

2. Convert the text to the conll format. The columns you create automatically are (**at least**) the following: token number in sentence, word, lemma, POS. You can add more columns with more information. You can either write your own code or use one of the two implementations written by Human Language Technology or Text Mining students, provided in the coreference package.

3. Find the coreference chains in the text. **You can limit yourself to the fragment printed above (i.e. *Chinese scientists identified the new virus strain last week.* would be the last sentence included).**

4. Add a column with coreference information for representing the coreference chains you found. The column should be in the format of the SemEval Shared Task 2010. It is explained in the conll video, but you can also read the paper that describes the shared task (`https://dl.acm.org/doi/10.5555/1859664.1859665`) and check the website `http://stel3.ub.edu/semeval2010-coref/`.[1] You can limit your annotations to the fragment **above** (ending with *Chinese scientists identified the new virus strain last week*) and only need to indicate chains (i.e. you may leave singletons out in this exercise). If you do decide to include singletons, then you should mark all of them. Also note that in the next question (1.1.5), singletons are specifically asked for and you should include them there.

   The resulting format should look like this (tab separated, with more columns if you decide to add more information):

1	The	the	DET	(1
2	dog	dog	NN	1)
3	that	that	WDT	(1)
4	I	I	PRP	(3)
5	saw	see	VBD	_
6	in	in	IN	_
7	my	my	PRP$	(3)
8	house	house	NN	_

5. List the mentions, singletons, anaphoras, cataphoras and bridging anaphora, if there are any, that occur in the fragment provided **below**. You can color code the mentions, if you find that more efficient. Three mentions have already been color-coded. If you have doubts about certain examples or phenomena, provide comments explaining what you doubt about and why.

   Please indicate how much time you spent on these 5 questions.

   Fragment for Question 5.:

   A Chinese tourist was found to be infected with the new strain of coronavirus when she arrived in Thailand, is being treated in hospital and is expected to be discharged in a few days, Public Health Minister Anutin Charnvirakul said on Monday.

   Mr Anutin said the 61-year-old woman was recovering at Bamrasnaradura Infectious Diseases Institute in Nonthaburi province.

[1] The video suggests indicating intermediate terms of a mention (when it consists of three or more terms, but the number without brackets. The SemEval description leaves those blank. Either representation is fine, as long as you use it consistently and it is clear what the coreference chains are.

> She now had no fever or any respiratory symptoms. If doctors gave her a clearance she would be allowed to go home in a few days, said Mr Anutin.
>
> Sixteen other people who were close to the woman on the same flight were examined, and the results were negative, he said.
>
> Mr Anutin said 59 people in China have been confirmed infected with the new strain of the coronavirus, which has been linked to a sudden outbreak of pneumonia in central China. One of them died. All had attended big markets selling animals and seafood in Wuhan city. They were either workers or buyers. There had not been any human-to-human transmission of the virus.
>
> The ill Chinese woman was the first person detected with the virus outside China.

1.2 Towards designing a system

In this exercise, we will walk through the steps of (1) analyzing phenomena, (2) identifying information that can be used for addressing it computationally and (3) designing a system.

Please note that we do not expect completely correct answers for everything at this point. If your answer shows you have thought about it carefully, it is good for now.

1.2.1 Analyzing the phenomenon

We start by analyzing the phenomenon. For the questions in this section, you should not use any material (not from class nor external). Just base your answers on your ability of understanding text in English and your current knowledge, providing explanations in your own words. If you do happen to remember an example or insight from the material, this is fine: you can of course make use of what you remember. Just do not go out of your way looking for explanations.

1. There are three mentioned highlighted in the text fragment of Exercise 1.1, *Mr Anutin*, *She* and *them*. In this exercise, you will reflect on how you can determine what the antecedent is in two steps. The first has been done for you as an example.

    - List their antecedent.
    - Explain how you know (based on your ability of reading this article) this is the correct antecedent, or why you think this is the most likely one. You may have more than one reason. Explain each reason in one or two sentences.

    Example answer for *Mr Anutin*:

    Antecedent: Public Health Minister Anutin Charnvirakul.
    Reason 1: The name Anutin matches.
    Reason 2: "Mr." indicates the antecedent is a male person and the only other person introduced so far is female.

2. Do the same for the following highlighted expressions: list the correct (or most likely) antecedent and explain with one sentence per reason why you selected this. Note that this may involve explaining why other candidate expressions can not be the antecedent. The sentences should be read independently.

   (a) Kim and his brother bet that Kim could take a picture of himself while doing a handstand.

   (b) Kim and his brother bet that Kim could take a picture of him while doing a handstand.

   (c) I went with John to the market yesterday and then took a walk in the forest with Mark. He was in a cheerful mood.

   (d) John met Mark for dinner. He felt like having company.

   (e) Kim was stopped by a policeman. He was not pleased.

   (f) John invited Mark for dinner. He knew he was broke.

   (g) John invited Mark for dinner. He knew he could please him with his latest recipe for lava cake.

   (h) John has a bike, but no car. It is red.

   (i) He brought flowers and chocolates. They tasted great.

   (j) She poured water from the pitcher into the cup until it was full.

   (k) She poured water from the pitcher into the cup until it was empty.

3. Provide a classification of the reasons you provided. You can list the type of reasons and indicate to which sentences they apply. Reasons may include:

   - Syntactically impossible (i.e. the form should have been different for an alternative interpretation)
   - Semantically impossible (e.g. alternative candidates have a different gender, number)
   - World knowledge (i.e. given the way the world works, this is a more logical or even only possible interpretation)
   - Structure (i.e. due to the chosen sentence structure/order in which elements are presented, this is the first interpretation that comes to mind)
   - Other reason (please specify)

1.2.2 Towards system design

In the next step, we start thinking in terms of system design. We start by thinking of ways we could feed the observations made in our analysis above to a computer based on our own knowledge. Then we look at some early approaches and see if we can make this more concrete. In this exercise, we are still in the exploration stage. Your answers thus need not be perfect.

(a) **Before looking at the material on coreference analysis** For each of the reasons (syntactic, semantic, world knowledge, structure, other) you identified in the previous exercise, try to imagine what would be needed to provide the necessary information to a computer program for coreference resolution, if this is possible at all. Think about it carefully and possibly creatively, but do not break your head over it. You do not need to do any reading for this component. Base your answer on what you know.

(b) **After looking at the material on coreference analysis** Based on the approaches you have seen so far, provide a description of what information you would feed to a computer program if you were to build a coreference system. You can limit yourself to the material included in Step 5 of the coreference component in Module 2 (specified in overview and planning). You do not need to dive into other approaches at this point.

(c) Propose a set of features that a supervised machine learning system could use to determine whether two mentions are coreferent. Explain for each feature (in 1-3 sentences), why it would be helpful.

Please let us know how much time you spent on this exercise.

1.3 Rounding up system design

This exercise starts with reading the following article from the beginning to Section 3.3 (included): H. Lee et al (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 38:4, pages 885-916. `https://www.aclweb.org/anthology/J13-4004.pdf`.

Do not worry too much if you do not understand all the details: if you can summarize each 'sieve' (the main steps in the algorithm explained), it is enough.

1. Following the order in which the *Passes* in Section 3.3 are presented, explain briefly what the sieves capture. Note that:

   - Explanations can be short (1-3 sentences) and should come with an example. You should use your own words and examples (i.e. do not copy-paste from the paper!).
   - Your overview need not necessarily be one explanation per sieve. If it is easier for you to explain or come up with examples:
     - You can group sieves that do more or less the same thing together in one explanation.
     - You may want to split some Sieves that capture multiple phenomena up in multiple explanations.

   Example:

   Sieve 1: Speaker Identification: In direct quoted speech, first person pronouns (*I, me,* *my* correspond to the speaker and second person pronouns to the addressee. For instance, in *Kim said to his Brother: "I will help you later"*, *Kim*, the speaker, corefers with *I* and *his brother* corefers with the addressee *you*.

You can use your own words (you need not need to familiarize yourself with new linguistic terminology if you can also explain using every day language.

2. For each of the examples in Exercise 1.2.1, Question 2, indicate which of the sieves would capture them, if any. You can add the sentence identifiers (a,...,k) to your explanations of Question 1 above and add a line that indicates which phenomena would not be treated correctly by the sieves, or provide an overview in a separate table (which ever you find more convenient).

3. Revise the feature set you proposed in Exercise 1.2.2 (c). Provide brief explanations (1-3 sentences) of why new features would be helpful. If you revise features, also explain what the revision is based on.

2 Distributional Semantics

In this component, you will explore distributional semantic models. We will first dive into their capability of capturing meaning, or to be precise, to identify which words have similar meaning and which do not. Since this form of evaluation aims to measure how well distributional semantic models capture semantics, we call this *intrinsic* evaluation.

The second part of this exercise investigates the impact of using a distributional semantic model to create features for an NLP task. An evaluation that investigates how suitable a model is for applications (or in general to be used as part of other systems) is called *extrinsic* evaluation.

2.1 Measuring Similarity: Intrinsic Evaluation

In this exercise, you will load a pretrained distributional semantic model, explore what it can do and run a standard intrinsic evaluation on it. You will then inspect the results and reflect on the outcome of the model.

You will work with the jupyter notebook Exercise2.1_intrinsic_evaluation.ipynb which is included in the materials provided as part of the assignment (links can be found on the assignment page and the Module page). Please note:

- The notebook makes use of the SimLex data provided under evaluation/. Paths are defined in the notebook. If you use a different local structure, please adapt the paths accordingly.
- You will need to download pretrained word embeddings. The code in the notebook assumes you download them from Google and place them under models/ in your working folder (the notebook provides pointers to where to download them from). You may use other embeddings as well.

- Loading the model and extracting the 'most similar terms' can take a bit of time

Exercise:

1. Walk through the jupyter notebook and follow the instructions. Once you have loaded the model, explore what it can do using the 'most_similar' and 'similarity' functions.

2. The notebook runs an experiment comparing the similarity scores assigned by the model you loaded to human judgments on the SimLex-999 set:

   Hill, F., Reichart, R., Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4), 665-695. `https://www.mitpressjournals.org/doi/pdfplus/10.1162/COLI_a_00237`

   If you have the feeling you understand enough of the dataset based on the lecture, you do not need to read the paper, but feel free to check it out otherwise.

   Report the outcome of the overall experiment. Mention which model you used, briefly explain what Spearman Rho measures (look that up if you are not familiar with it) and report the score.

3. Inspect the rankings made by humans and by the distributional semantic model. You will find the human rankings ordered and with scores in the file evaluation/SimLex-999.ordered.pairs.csv. The notebook should have created an output file with an ordered ranking by the model you used. You can either inspect the files manually or write a small program that identifies clear differences.

   (a) Identify a couple of cases where humans rated a pair to be much more similar than the model and cases where it was the other way around (the model ranked a pair higher than humans). Both cases are considered "mistakes" by the model (they have a negative impact on the Spearman rho and thus on how well the model 'scores'.

   (b) Reflect on the mismatches you observed between people and the model. Can you think of any reason why the model ranked them differently from people? And is the model always wrong? Your answer should be 1/2 - 1 page.

How long did it take you to complete this exercise?

2.2 Machine Learning: traditional features and embeddings

In this component, you will explore the impact of individual features on a machine learning system. After a small study on the impact of individual "traditional" features, you will replace the one-hot representation of tokens by word embeddings. The code for this assignment is provided in the folder NLP_tech_Assignment2.2.zip on Canvas. **Note: loading the distributional semantic model and training the system with word embeddings may take some time (running the full set of experiments without feature ablation took about 50 minutes).**

Open the notebook `Embeddings_in_ML_Assignment2_2.2.ipynb`. The notebook walks you through the following steps:

- Creating a basic system for NERC (with just tokens as features)
- Running a basic evaluation (providing a confusion matrix and calculating precision, recall, f-score)
- Adding additional features
- Using the code to select a subset of features (you will need this for running a feature ablation study)
- Illustrating what one-hot encoding looks like
- Integrating pretrained word embeddings as feature representations in a classifier (first only the token itself then combined with the previous token)
- Combining pretrained word embeddings and traditional features

Use the notebook to carry out the following experiments and write a brief report on this. The steps below describe what should be included in the report (i.e. note that a general introduction, related work section, complete system architecture and method section are **not** necessary).

1. **A NERC classifier** Use the notebook to train classifiers for named entity recognition up to the more elaborate system in Step 3. Provide a brief description (with motivation) of the features used in this system. (Note: you may include other features if you want, but this is not required).
2. **Ablation Analysis** Move on to Step 4. In this step, you will carry out a feature ablation analysis. You will investigate what the impact of each individual feature included in the more elaborate system (Step 3) is by systematically testing all feature combinations. The cell in Step 4 illustrates how you can train a system with a subset of available features. You can adopt the list of selected features and rerun the cell manually or write a script that systematically goes through various feature combinations. Provide an overview of the results in a table and a description of your main observations.
3. **Word Embeddings** Run Part 2 of the notebook, where word embeddings are used to represent tokens. Make sure that you have a distributional semantic model stored on your machine and adapt the path to the model if necessary (the current implementation assumes you have the Google 300-dimensional news word embeddings created with negative sampling stored in the folder models/). Report on the outcome of the experiments. Make sure to specify which pretrained model you used.
4. Write a short conclusion commenting on the impact of individual features and of using word embeddings as features.

Acknowledgements

Several exercises in the coreference assignment are taken from an assignment developed by dr. Roser Morante. This applies to all questions in 1.1, except for the change that we are providing code for generating the conll files.