LIN3011/LIN3012 Data-Driven NLP

Final Projects

Instructions

Choose **one** of the projects described below. Each project should be delivered through the VLE upload area in the form of a write-up and associated code, by the deadline indicated on the VLE.

Projects need to be done individually.

The deadline for submission is Februry 17, 2019.

Layout

Your assignment **must** incorporate sections on the following:

1. An introductory section, giving a problem definition and an appropriate motivation.
2. A brief review of relevant literature, showing that you have familiarised yourself with related work that seeks to solve this, or similar, problems.
3. Data and Methodology. This is the most important section. It must include:
    a. A full description of the data used, including any train/dev/test splits and how they were determined.
    b. A description of the methods used, with details of architecture, hyperparameters etc.
4. Results, presented clearly using tables, figures or any other appropriate media.
    a. Where possible, compare your results directly with existing state of the art.
    b. If you have used a baseline, be sure to compare your main models with the baselines.
    c. Where possible, in addition to figures on performance, perform a short qualitative analysis, showing examples of outputs accompanied by a discussion of where things went wrong (and where they went well).
5. Discussion and conclusions

Code and data

- Ony submit data if it isn't a standard release, Ie. If you created it yourself from scratch or by modifying an existing dataset. If you just used an existing dataset, you don't need to submit it, but include a reference to the relevant work describing it.
- Code can be written in a programming language of your choice. Ensure that code you submit includes:
    - Brief but clear documentation
    - A README file that allows one to run it
- Code and data should be submitted together as a zipped archive.

Length

Projects should be between 8 and 10 pages, double-spaced, 12pt font.

The above length limit does **not** include bibliography.

Assessment criteria

Projects will be assessed based on the following criteria:

- **Methodology (50%)**: Is your method appropriate for the chosen topic? Did you train and test your models properly and are you reporting appropriate baselines for comparison?
- **Presentation and write-up (25%):** Does your write-up incorporate an appropriate literature review, data/methods presentation and discussion? Do you present your results in a transparent form, using tables, graphs etc? If you conducted some form of supervised learning, do you describe your features appropriately?
- **Coding or data collection effort (25%):** If you submit code, is it well-commented? Does it actually do what it says on the tin? If you've collected additional data, does your report describe it properly and is it clear what you've collected and why?

Note that I do not accept project write-ups which are basically just code documentation. This is not primarily a coding project. Your aims are scientific – your code is a tool to achieve those aims. Your write-up is a scientific report, not a manual.

Project 1: Blog gender identification

This project is concerned with the use of statistical models or classifiers to identify the gender of authors of blogs. Work in this area has identified a number of interesting variables in people's use of language that helps to identify them. Examples include their use of function words, the hapax legomena in their text, etc.

Your aim in this project is therefore to identify the linguistic markers of gender, justifying your choice of features with reference to the literature, and applying a machine-learning methodology to conduct your experiments.

The steps involved are as follows:

1. Find blog texts written by different authors of different genders. You can use the corpus linked below, or a sample of it.
2. Identify gender-relevant features.
3. Build and train a model which, given an input text by one of the authors, extracts its features and classifies it according to the most likely author gender. Note that it is possible to view this as a classification task. It is up to you to choose the classification algorithm to deploy. You should, however, compare your results to an appropriate baseline.
4. Evaluate the model using test data or using a cross-validation design.

Useful resources:

1. The Blog Authorship Corpus, constructed by Moshe Koppel, is a very large corpus of blog posts by multiple authors, with some demographic variables about authors available (e.g. age, gender and astrological sign)
2. An oldish paper to get you started (but you'll need to go further than this): J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of age and gender on blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*

Project 2: Part of speech tagging

The task for this project is to design and implement a probabilistic part of speech tagger for Maltese.

There currently exists a POS-tagged corpus of Maltese, divided into several sections. This corpus is called *Korpus Malti v3.*0. The Corpus itself can be browsed online, or downloaded. It is divided into several sections. For this project, you are advised to **use the "newspaper" section** of the corpus. (See below for details).

The corpus texts are in a vertical format, as shown in the following example:

Fl-	PREP-DEF	fi	null
aħħar	NOUN	aħħar	null
mill-	PREP-DEF	minn	null
aħħar	NOUN	aħħar	null
,	X-PUN	,	null
kitbu	VERB	kiteb	k-t-b
li	COMP	li	null
l-	DEF	il-	null
għan	NOUN	għan	null
ta'	GEN	ta'	null
CMM	X-ABV	CMM	null
huwa	PRON-PERS	huwa	null
li	COMP	li	null
joffri	VERB	offra	null
attivitajiet	NOUN	attività	null

In the above, the first column is the actual word used. The second column is the POS Tag, the third is the lemma (the morphological base form, for example, in the case of *attivitajiet*, which is a plural noun, the base form is the singular; similarly, for *kitbu*, the base form is *kiteb*). The final column is the root (which is null for all words except those of Arabic origin, which have a consonantal root, as in the case of *kitbu* above, whose root is *k-t-b*).

You should treat this corpus as your training and test data, i..e you will be developing a new POS Tagger and comparing the outcomes against the corpus.

Here is what you minimally need to do:

1. Train your tagger.
2. Test the tagger on held-out test data.
3. Evaluate the results. This should also include an error analysis, i.e. a discussion of where the tagger goes wrong.

You are free to choose any technique for training your POS tagger. The corpus itself has been tagged using a tagger based on support vector machines, called the SVMTool. It reaches an accuracy of around 96%. Can you come up with a method to beat that?

NOTE: If you wish to do POS Tagging for a different language, other than Maltese (but please, not English, because that's boring), please get in touch with me, and let's discuss it.

Useful resources:

- The Maltese Corpus is available on the Maltese Language Resource Server: http://mlrs.research.um.edu.mt. If you visit the "Downloads" section of this page, you will be able to download individual sections of the corpus, including the newspaper section.
- The Part of Speech Tagset developed for Maltese and used in this corpus is described here: http://mlrs.research.um.edu.mt/resources/malti03/tagset30.html
- There is a huge literature on POS Tagging. Jurafsky and Martin's textbook offers a useful entry-point into the literature.
- The SVMTool used to train the corpus can be found here: http://www.lsi.upc.es/~nlp/SVMTool/

Project 3: Stylistically controlled sentence generation

In this project, you will be generating texts based on stylistic features. The dataset used was developed by Ficler and Goldberg (2017). It is a set of movie reviews harvested from the website rottentomatoes.com .

Ficler and Goldberg (2017) are interested in controlling linguistic style in automatically generated text. Specifically, they are interested in modelling the features of a sentence in a review that determine, among other things:

1. Whether it is professionally written by a critic, or by an amateur;
2. Which aspect of a movie (e.g. plot, acting) the sentence describes;
3. Whether the sentence is descriptive;
4. Whether the sentence is written in a personal ("I like this…") or objective ("This movie is…") style;
5. The overall sentiment of the review, based on the rating given by the critic.

Your task in this project is to design and implement a model that generates sentences. The model should be defined as follows:

- INPUT: A set of parameter values according to the five features above
- OUTPUT: A sentence in English that reflects the parameter values.

In this project, you are **strongly encouraged** to also experiment with additional stylistic features, over and above what Ficler and Goldberg used.

Useful resources:

The paper from which this data is obtained:

- Ficler, J., & Goldberg, Y. (2017). Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation* (pp. 94–104). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1707.02633
- **NB:** THe supplementary material contains hints about how to annotate each sentence automatically.
- The data is linked from the VLE. It is provided to you with the kind permission of Jessica Ficler. **Do NOT distribute this data.**

Widely-used language modelling toolkits include the following:

- The SRI language modeling toolkit: http://www.speech.sri.com/projects/srilm/
- The CMU-Cambridge language modeling toolkit: http://www.speech.cs.cmu.edu/SLM/toolkit.html
- NLTK's built-in libraries for langauge modelling
- There are plenty of libraries for training and testing recurrent neural nets, if you decde to implement a neural model. You get additional credit for modelling this using a neural network language model, as Ficler and Goldberg in fact did.

Project 4: Multilingual EMOJI prediction

Emojis are part and parcel of people's interactions on social media, where they convey everything from emotion to content that can be concisely expressed in a non-purely linguistic form.

From a linguistic perspective, it seems reasonable to predict that the use of an emoji (say, a heart) in the context of a longer message, such as a tweet, is somehow connected with the content of that message.

For this project, you will be experimenting with a task that was conducted as a shared task at the SemEval 2018 conference. The setup of this task was simple:

- Data: 500k tweets in English, and 100k tweets in Spanish, each of which contains exactly one emoji.
- THe emojis fall into one of exactly 20 possible classes, which function as the labels in this task.
- The aim is, given a tweet, to predict which emoji is the correct one.

Useful resources:

A full description of the Shared task, including a summary of the results by various participating systems, can be found here:
https://competitions.codalab.org/competitions/17344#learn_the_details-data

Since this is a shared task, the data, including the test data, is all available on the above URL.

Note that the task organisers also provide an evaluation script. Feel free to use this, but be sure to understand what it does, and include details of this, and any other evaluation method you decide to use, in your report.

Useful literature:

- The shared task results are described in this paper by Barbieri et al (2018)
- Useful background on emoji prediction can be found in this paper