Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Micha de Rijk		
Název práce	Codenames: a practical application for modelling word association		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	Mgr. Martin Popel, Ph.D.	**Role**	oponent
Pracoviště	ÚFAL MFF UK		

Text posudku:

The thesis studies several word-association measures within the task of artificial intelligence for a single-player version of the Codenames game. I find this thesis topic challenging and very interesting from the theoretical point of view. At the same time, there is also a nice practical achievement – a web-based Codenames game, which I found quite enjoyable to play. Below, I review three aspects of the thesis: the application (game), the models used and the text of the thesis.

Application

I was satisfied with the implementation of the game, which is simple, but functional and easy to understand and use. I have just a single suggestion for improvement in this aspect – it would be nice to reveal which hint was targeting which words (cards) using which type of association.[1] While showing this information during the game could bias the results (as noted by the author), I don't see any reasons preventing to show this after the end of the game.

Models and their evaluation

The author has chosen a methodologically sound approach, starting with baseline (random clicking) models, continuing with simple models, which were gradually improved by weighting, thresholding and simple ensembling. In general, I find the number of experiments and their evaluation sufficient for a Master thesis. I appreciate especially a thorough discussion of the results.

- The TopN-Mutual model is not sufficiently described. There can be multiple hints predicted by the two models and it is not clear which hint is chosen. Obviously, the hints' scores (or just ranks) have to be taken into account in order to prefer the better hints, but the description "we let both models predict hints, until one of the models gives a hint that the other model has also predicted" is not sufficient.

[1] In case of sentence/dependency collocations, one could imagine showing also few example sentences from the training data which contributed to the high PMI score.

- I miss an ensemble combining several models (e.g. dependency and embeddings) for the same hint, where e.g. the score of a given hint and each target word is computed as the maximum (or sum or another function) of the weighted similarity scores of the individual models. I appreciate this direction is reflected in Section 6.1 Future Work.

Thesis text

The thesis is written in English, it is well organized and mostly easy to follow. My comments are:

- The last paragraph on page 9 is redundant with the previous paragraph.
- The hint filter is reported to filter out "almost all morphologically related words", but what about e.g. *shoes* and *horseshoe*? Neither one is parts of the other and the relative Levenshtein distance is 6/9 = 66%. One could try lemmatizing/stemming the words before checking substrings.
- Section 3.6 says "Similarly, hooray is likely related to day through the word birthday." I don't understand this explanation of high dependency-level and sentence-level similarity scores between words *hooray* and *day*. *Birthday* (unlike *ice cream*) is written as a single word. A sentence containing *birthday* and *hooray* but not *day* does not increase the PMI(day, hooray).
- It is not clear how precision, recall and f-score are computed. The reader could be reminded with the well-known formulas (which are not shown at all, unlike the PMI formula, which is shown three times), but this is not my point. My confusion stems from Section 5.1, which says "the false negatives are the player's cards that they did not click at the end of the game" and "using precision, recall and f-score on the decision level instead of win and loss rate at the game level". So game-level false negatives were not used for computing the decision-level precision, recall and f-score? Or was the same number of false negatives used for all player's turns (decisions) within a single game?
- Table 5.10 states some statistics (number of games and decisions) about the experiments with human players, but some interesting statistics are missing, e.g. the number of unique players. Also, Table 5.10 should be moved before Table 5.2, or at least referenced there. Ideally, confidence intervals should be computed for all the reported results and significance should be reported for any comparison (e.g. Enemy>Neutral in two models in Figure 5.2).

There are several grammar/style errors/typos, but also some factual errors/typos, e.g.:

- page 33: "0.389, **0.39** and 0.362" → "0.389, **0.339** and 0.362"
- page 48: "The TopN word embeddings model uses the Top3, Top2 and Top1 **dependency** models" → "...and Top1 **word embeddings** models"

Questions

- How are the f-scores exactly computed? What are the false negatives? (cf. my comment above)
- Is there any rule preventing the model to give the same hint again (for the same (sub)set of target cards, if they were not guessed in previous turns)? A more general question: does the history (of hints and clicked cards from previous turns) influence the currently suggested hint?
- Was the same set of players involved in the Top1, Top2 and Top3 experiments as in the TopN experiments? May the lower TopN results be caused by "hiring" new/inexperienced players?

I recommend the thesis to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 28. 1. 2020

Podpis: