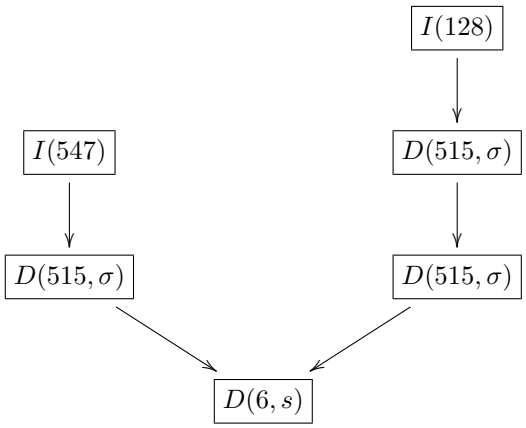
Response to the "My Tailor is Rich" Challenge

Yannis Haralambous* Philippe Lenca†

May 28, 2018

1 Global Architecture

We used a double-input single-output feed-forward NN, as follows:



where $I(547)$ and $I(128)$ are input layers of 547 and 128 dimensions respectively, $D(515,\sigma)$ is a dense 515-neurons-output feed-forward layer with sigmoid activation, and $D(6,s)$ is a dense 6-neurons-output feed-forward layer with softmax activation. We used 42 epochs and a batch size of 32.

The network was implemented in functional `keras`, on

1. MacOS X 10.12.6 on a MacBook Pro, 4 × 2,9 Ghz Intel Core i7, with 16 GB of RAM and a Radeon Pro 455 2048 Mo VRAM and Intel HD Graphics 530 1536 Mo VRAM GPU,
2. a Debian 9 on Intel(R) Core(TM) i7-3930K CPU with 12 × 3.20 GHz, 32 GB of RAM and a NVIDIA GF119 [GeForce GT 610] GPU.

2 Linguistic Indicators

We have analyzed both corpora (training and prediction) according to dependency syntax using the Python `spacy` package. We have also processed both corpora with TAALES (= *Tool for the Automatic Analysis of Lexical Sophistication*) software by Kristopher Kyle[1].

*IMT Atlantique & Lab-STICC, `yannis.haralambous@imt-atlantique.fr`
†IMT Atlantique & Lab-STICC, `philippe.lenca@imt-atlantique.fr`
[1]`http://www.kristopherkyle.com/taales.html`

2.1 Left Branch

The 547 linguistic indicators we used are defined as follows:

1. the 58 values provided by the challenge;
2. 1 value representing badly spaced punctuation (no space after period or comma, etc.);
3. 2 values for number agreement errors: (a) whenever the subject of the sentence is in plural and the verb in singular, (b) whenever the subject is in singular (and there is no conjunction) and the verb is in plural;
4. 1 value representing errors in the use of prepositions in phrasal verbs (verbs used with wrong prepositions, or without preposition when a preposition is necessary);
5. 1 value representing the sum of inverse quadratic frequencies of words in the Simple English Wikipedia;
6. 547 values provided by TAALES software.

The resources we used:

- Simple English Wikipedia;
- a list of English phrasal verbs;
- TAALES software.

2.2 Right Branch

We have processed syntax trees of corpora with `graph2vec` and obtained a 128-dimensional embedding of the texts (both training and prediction). We then applied a feed-forward layer to perform classification using this embedding.