Escuela de Ciencias Informáticas 2019

Curso M1

Procesamiento del lenguaje natural mediante redes neuronales

Trabajo practico: replicación de los resultados del paper

"Annotation Artifacts in Natural Language Inference Data"

Facundo Emmanuel Messulam[*] Ramiro Hernán Gatti[**]

10 de agosto de 2019

[*]Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Licenciatura en Ciencias de la Computación

[**]Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática, CONICET-UNER

Preliminar

Uno de los datases mas famosos para estudiar la inferencia dentro del Lenguaje Natural es el 'The Stanford Natural Language Inference (SNLI) Corpus'. El mismo consiste de una serie frases que corresponden a una premisa (A) y su respectiva hipótesis (B). Dada A se deriva B que puede ser una implicación (*'entailment'*), contradictorio (*'contradiction'*) o neutral (*'neutral'*) respecto de A. En el trabajo de Jouling y otros [1] se muestra que un sesgo en el dataset permite clasificar las hipótesis sin necesidad de la premisa utilizando el clasificador de Facebook *'fasttext'*.

En el presente trabajo se busca replicar y mejorar el resultado obtenido por Jouling y otros [1]. Una primera implementación utilizando *'fasttext'* con los parámetros por defecto, permite obtener una clasificación con valor predictivo positivo (VPP) y tasa de verdaderos positivos (TVP) del 64 % en los datos de validación. La cual se encuentra marcadamente por encima de la tasa teórica de clasificación al azar para 3 clases de 33 %, aproximadamente el resultado esperado para un dataset que no tiene ningún tipo de información sobre la premisa.

Primeras optimizaciones

En una primera optimización se eligió transformar las palabras a vectores de 325 dimensiones y tener en cuenta la información de las palabras circundantes por medio de bigrmas [2]). De esta forma se logra obtener valores de VPP y TVP cercanos 66 % en los datos de validación.

Optimización por etiquetación sintáctica

Hasta el momento se utilizó simplemente la oración para clasificar. Sin embargo, al tomar como entrada del clasificador la oración y el árbol binario con la separación sintáctica:

```
The sisters are hugging goodbye while holding to go packages after
just eating lunch. ( ( The sisters ) ( ( are ( ( hugging goodbye )
( while ( holding ( to ( ( go packages ) ( after ( just ( eating lunch
) ) ) ) ) ) ) ) ) . ) )
```

De esta forma se utiliza mayor información disponible dentro del dataset de entrenamiento y se obtiene valores de VPP y TVP de 67 %, mayor que con solo el árbol o la oración.
Valga aclarar que si bien se tiene una separación sintáctica, que de hecho contiene la información de que es cada nodo en el árbol:

```
(ROOT (S (NP (DT A) (NN person)) (VP (VBZ is) (VP (VBG training) (NP
(PRP\$ his) (NN horse)) (PP (IN for) (NP (DT a) (NN competition)))))
(. .)))
```

Usar esta versión empeora el resultado.

Explicación teórica

Supongamos que se tienen cuatro oraciones, una premisa y tres hipotesis. Ahora bien, sabemos por la creación del SNLI[3] que cada hipótesis esta relacionada con la premisa por un contexto. Si bien no se puede afirmar que el contexto sea enteramente deducible de todas las hipótesis, se puede afirmar que el contexto influenció de forma casi única a los contenidos de la oración.

Referencias

[1] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov. *Bag of Tricks for Efficient Text Classification*. Association for Computational Linguistics, 2017

[2] Sida WangandChristopher D. Manning. *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification*. Stanford University, Department of Computer Science, 2018

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015.*A large annotated corpus for learning natural language inference*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).