Escuela de las 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‡

August 9, 2019

*https://www.aclweb.org/anthology/N18-2017

†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

El paper propone el uso de "fasttext" el clasificador lineal de Facebook[1]. Una primera implementación da una clasificación correcta del 64%. Que permite empezar ya a apreciar una marcada diferencia con el 33% que "deberia" ser para un dataset que no tiene ningun tipo de informacion en la hipotesis.

Primeras optimizaciones

Tomando vectores de largo 325 (buscado por descenso del gradiente) y tomando información de las palabras circundantes (con bigramas[2]), se logra un puntaje de 66%.

Optimización por etiquetación sintáctica

Hasta ahora usamos simplemente la oración para clasificar. Sin embargo si tomamos la oración y le agregamos un arbol binario con la separación sintatica:

```
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
) ) ) ) ) ) ) ) ) . ) )
```

Se obtiene un puntaje 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 arbol:

```
(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 puntaje.

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 influeció de forma casi unica a los contenidos de la oración.

References

[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).