Universitat Oberta de Catalunya

Patricia Lázaro Tello
plazarotello@uoc.edu

PEC5: Revisión del estado del arte

Efficient Estimation of Word Repesentations in Vector Space[*]

Mikolov, Chen, Corrado y Dean

Una tarea común de los procesadores de lenguaje natural (NLP) es la predicción, también llamada estimación, de la siguiente palabra en una oración. La mayoría de NLPs actuales utilizan técnicas que tratan cada palabra como una unidad atómica, como los modelos de N-gramas.

Una técnica en alza en los últimos tiempos es el tratamiento de palabras como vectores continuos; es decir, cada palabra es transformada a un *array* de números. Su principal ventaja es la obtención de medidas de similitud entre palabras, tanto sintáctica como semánticamente. Estas similitudes y la representación vectorial permiten el uso de operaciones algebraicas simples para obtener relaciones entre palabras.

Esta técnica es utilizada mayormente por redes neuronales (*neural network language model* o NNLM), y se ha observado que produce mejores resultados que la técnica tradicional de tratar palabras como unidades atómicas, llegando a simplificar incluso los NLP.

Sin embargo, las redes neuronales presentan también algunos inconvenientes, como puede ser la **complejidad computacional** que impide en la práctica entrenar los modelos con más de 50-100 características en los vectores de palabras y más de unos cientos de millones de palabras:

En este *paper* se proponen 2 modelos de arquitectura que minimizan la complejidad computacional y permiten, por tanto, aumentar las dimensiones de los vectores y el volumen de datos de entrenamiento. Estos modelos poseen una *accuracy* superior a la de las redes neuronales.

- **Modelo *Bag-of-Words* continuo**: para predecir la palabra en una posición *t*, se utilizan las palabras adyacentes en una ventana *n*, o dicho de otra manera, se utilizan *{t-n, t-n+1, ..., t-1, t+1, ..., t+n-1, t+n}* palabras para predecir *t*. La posición de las palabras en la oración no es importante.
- **Modelo *Skip-gram* continuo**: su función es predecir las palabras alrededor de una palabra en una posición *t*. Se predicen las palabras adyacentes en una ventana *n*, o dicho de otra manera, se utiliza *t* para predecir *{t-n, t-n+1, ..., t-1, t+1, ..., t+n-1, t+n}* palabras.

Para comparar la calidad de los modelos propuestos respecto a la calidad de los modelos ya existentes se ha utilizado una tabla de relaciones semánticas y sintácticas aprendidas y una tabla comparativa de *accuracy* semántica y sintáctica, así como de tiempos de entrenamiento. También se han utilizado los resultados del *Microsoft Research Sentence Completion Challenge* (58.9 % de *accuracy* contra el 55.4 % obtenido hasta el momento).

Aunque las redes neuronales ofrezcan mejores representaciones de las palabras, el entrenamiento de modelos simples como los propuestos con muchos más datos presenta **mejores resultados en *accuracy* y coste computacional**.

[*]Mikolov, Tomas; Chen, Kai; Corrado, Greg y Dean, Jeffrey, 2013. *Efficient Estimation of Word Representations in Vector Space*. Disponible desde arXiv: 1301.3781.