Evaluación del Módulo 4

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Febrero 2020

1. **Razone si es importante el orden de las palabras en traducción automática (2 puntos).**

   El orden de las palabras en traducción automática es muy importante, ya que el orden de las palabras dentro de los fragmentos de texto a traducir que se van leyendo, determina el resultado de la traducción. Si se modifica el orden de las palabras dentro de estos fragmentos, se pueden provocar incoherencias y cambios de sentido en las traducciones.

   Actualmente, es un desafío para la traducción automática, ya que como es bien sabido, el orden de las palabras en una frase difiere de lengua a lengua, y normalmente, el número de palabras en frases traducidas son diferentes.

   Podemos comprobar su alto grado de importancia dentro de las traducciones con un ejemplo,

   *L*a expresión española "viaje inesperado" se convertiría en inglés en "unexpected journey", donde podemos ver que se tiene que invertir el orden de nombre y adjetivo.

2. **¿En qué consisten los modelos de traducción basados en frases? (2 puntos).**

   Los modelos de traducción basados en frases son modelos de traducción que se basan en el alineamiento que se produce a nivel de frase o en unidades lingüísticas inferiores a la frase. Normalmente estos modelos generan diccionarios de pares de frases alineados.

   Con estos modelos se han intentado reducir las restricciones producidas por la traducción basada en palabras, traduciendo secuencias de palabras a secuencias de palabras.

   Los elementos a estudiar en este tipo de modelos son:

   - Encontrar la segmentación óptima en frases del texto a traducir.
   - El orden en el que se aplican los pares de frases traducidas.

Ejemplos

El modelo de traducción basado en frases más utilizado y de código abierto se llama Moses .

Moses es un sistema estadístico de traducción automática que permite entrenar automáticamente modelos de traducción basados en frases para cualquier pareja de idiomas. Todo lo que se necesita es una colección de textos traducidos o corpus paralelo. Una vez que se ha entrenado el modelo, un algoritmo de búsqueda eficiente encuentra rápidamente la traducción de mayor probabilidad entre el número exponencial de opciones disponibles.

Otro ejemplo es Stanford Phrasal . Phrasal es un sistema de traducción automática basado en frases estadísticas de última generación. En esencia, proporciona la misma funcionalidad que el núcleo de Moses. Las características distintivas con el modelo Moses es que proporciona una API fácil de usar para implementar nuevas características del modelo de decodificación, la capacidad de traducir utilizando frases que incluyen espacios y la extracción condicional de tablas de frases y modelos de reordenamiento léxico.

Otros ejemplos interesantes: Jane , Thot , Phramer , etc.

3. ¿Es cierto que la medida de evaluación BLEU es totalmente automática, es decir no requiere de ningún tipo de intervención humana, ni para la elaboración de un conjunto de referencia en el caso de que lo necesitase? (2 puntos).

No es cierto que la medida de evaluación automática BLEU sea totalmente automática. BLEU es una medida de las diferencias existentes entre una traducción automática y una o varias traducciones humanas de referencia de una misma frase de origen.

Como tiene ciertas deficiencias en la medida, se utiliza una precisión modificada con el fin de solventarlas; y en este instante es donde hay intervención humana. Pero también es su principal ventaja, se correlaciona muy bien con el criterio humano calculando el promedio de los errores de criterio de frases individuales de un corpus de prueba, en lugar de intentar averiguar el criterio humano exacto para cada frase.

4. ¿En qué consiste la desambiguación?

La desambiguación consiste en identificar el significado de una palabra en un determinado contexto dentro de un conjunto determinado de significados candidatos. Requiere de un diccionario para especificar los significados candidatos que deben ser desambiguados y un corpus de datos del lenguaje que donde también deben ser desambiguados (en algunos métodos se requiere un corpus de entrenamiento de ejemplos de lenguaje).

A continuación vemos un ejemplo de desambiguación.

La palabra "vela" puede tener dos significados: cilindro de cera o sebo, atravesado por una mecha que se prende para alumbrar; o pieza de lona o lienzo fuerte que, atada a las vergas, recibe el viento que impulsa la nave.

y en las oraciones:

- Puso dos velas a San Pancracio.
- Los egipcios fueron los primeros constructores de barcos de vela de los que se tiene noticia.

Para un ser humano, es evidente que en la primera frase se utilice la palabra "vela", como primer significado, y en la segunda frase, la palabra "vela" está siendo utilizada con el segundo.

El desarrollo de algoritmos para reproducir esta capacidad humana (desambiguar el significado) a menudo puede ser una tarea muy difícil. Se han investigado una gran variedad de técnicas, desde métodos basados en diccionarios que utilizan el conocimiento codificado en recursos léxicos, hasta métodos supervisados de aprendizaje automático en los que se capacita a un clasificador para cada palabra distinta en un corpus de ejemplos anotados manualmente por significado, hasta métodos no supervisados que agrupan las ocurrencias de las palabras. Pero de entre estos, los enfoques de aprendizaje supervisado han sido los algoritmos más exitosos hasta la fecha.

Un dato muy interesante es que existe una organización internacional Senseval dedicada a la evaluación de los sistemas de desambiguación mediante una competición que tiene como objetivo final comprobar la potencia y debilidad de los sistemas presentados. Esta serie de evaluaciones proporcionan un mecanismo para caracterizar en términos más precisos exactamente lo que es necesario para calcular el significado. También se centran en la efectividad, la eficiencia, el coste de producción (implementación y disponibilidad de recursos), etc.

Además, usan las típicas métricas de evaluación que se utilizan en recuperación de información,

- Precisión = (palabras desambiguadas correctamente) / (palabras desambiguadas)
- Recall = (palabras desambiguadas) / (palabras ambiguas)

5. Explique el concepto de ventana contextual y relaciónelo con el concepto de representación continua de palabras (word embeddings) (2 puntos).

Durante el proceso de desambiguación, se utiliza el entrenamiento con ventana textual. Que consiste en representar en un vector, cada término del corpus de entrenamiento, con el peso del término en el párrafo y con los pesos de los términos de la ventana contextual. Esta ventana contextual contiene palabras, con peso del término en el párrafo o con peso de los términos que constituyen la ventana contextual, circundantes al término a desambiguar.

El procedimiento del entrenamiento consiste primero en construir los vectores y calcular los pesos, anteriormente nombrados. Después se adaptan los algoritmos de aprendizaje, de realimentación por relevancia Rocchio o de redes neuronales Widrow-Hoff .

Por otra parte, el word embedding es el nombre de un conjunto de lenguajes de modelado y técnicas de aprendizaje en procesamiento del lenguaje natural (PLN) en dónde las palabras o frases del vocabulario son vinculadas a vectores de números reales. Conceptualmente implica el encaje matemático de un espacio con una dimensión por palabra a un espacio vectorial continuo con menos dimensiones. Este espacio se puede denominar ventana contextual.

Algunos ejemplos de word embeddings que circulan por Internet: Indra , word2vec , gensim , etc.