Capítulo 1

Procesamiento del Lenguaje Natural

El volumen de datos textuales digitalizados que se genera cada día es enorme (por ejemplo, la web, redes sociales, registros médicos, libros digitalizados). Por lo tanto, también crece la necesidad de traducir, analizar y gestionar esta avalancha de palabras y texto.

El procesamiento del lenguaje natural (PLN) es el campo que se encarga de diseñar métodos y algoritmos que toman como entrada o producen como salida datos de **lenguaje natural** no estructurado [Goldberg, 2017]. El PLN se centra en el diseño y análisis de algoritmos computacionales y representaciones para procesar el lenguaje humano [Eisenstein, 2018].

1.1. Tareas de PLN

Una tarea común de PLN es el Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés). Por ejemplo:

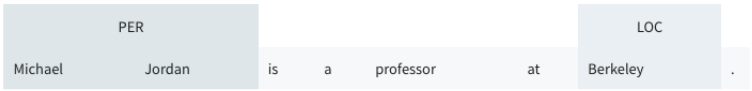

Figura 1.1: Reconocimiento de Entidades Nombradas

El lenguaje humano es altamente ambiguo, como en las frases: Çomí pizza con amigos", Çomí pizza con aceitunas.º Çomí pizza con un tenedor". Además, el lenguaje está en constante cambio y evolución, como ocurre con los hashtags en Twitter.

1.2. PLN y Lingüística Computacional

El procesamiento del lenguaje natural (PLN) desarrolla métodos para resolver problemas prácticos relacionados con el lenguaje [Johnson, 2014].

Algunos ejemplos son:

- Reconocimiento automático del habla.
- Traducción automática.
- Extracción de información de documentos.

La lingüística computacional (LC) estudia los procesos computacionales subyacentes al lenguaje (humano).

- ¿Cómo comprendemos el lenguaje?
- ¿Cómo producimos el lenguaje?
- ¿Cómo aprendemos el lenguaje?

El PLN y la LC utilizan métodos y modelos similares.

1.3. Diferencia entre PLN y LC

Aunque existe una superposición sustancial, hay una diferencia importante en el enfoque. La LC se centra en la lingüística respaldada por métodos computacionales (similar a la biología computacional o la astronomía computacional). En lingüística, el lenguaje es el objeto de estudio. El PLN se centra en resolver tareas bien definidas relacionadas con el lenguaje humano (como la traducción, la respuesta a consultas, las conversaciones). Si bien los conocimientos lingüísticos fundamentales pueden ser cruciales para realizar estas tareas, el éxito se mide en función de si y cómo se logra el objetivo (según una métrica de evaluación) [Eisenstein, 2018].

El procesamiento del lenguaje natural y la lingüística computacional están estrechamente relacionados y se superponen en muchos aspectos. Ambos campos utilizan métodos y modelos similares para abordar problemas relacionados con el lenguaje humano. Sin embargo, la diferencia principal radica en el enfoque: la lingüística computacional se centra en la lingüística respaldada por métodos computacionales, mientras que el procesamiento del lenguaje natural se centra en resolver tareas prácticas relacionadas con el lenguaje. Ambos campos son fundamentales para comprender y aprovechar el poder del lenguaje humano en la era digital.

1.4. Niveles de descripción lingüística

El campo de la **descripción lingüística** abarca diferentes niveles:

- **Fonética y fonología:** estudio de los sonidos del habla.
- **Morfología:** estudio de la estructura de las palabras.
- **Sintaxis:** estudio de la estructura de las oraciones.
- **Semántica:** estudio del significado de las palabras y oraciones.
- **Pragmática:** estudio del uso del lenguaje en el contexto.

El PLN puede abordar tareas en cada uno de estos niveles, pero a menudo se enfoca en niveles más altos de representación y comprensión.

1.5. Fonética

La fonética es la rama de la lingüística que se ocupa del estudio de los sonidos del lenguaje. Examina los órganos utilizados en la producción de sonidos, como la boca, la lengua, la garganta, la nariz, los labios y el paladar. Los sonidos del lenguaje se dividen en vocales y consonantes. Las vocales se producen con poca restricción del flujo de aire desde los pulmones, mientras que las consonantes implican alguna restricción o cierre en el tracto vocal [Johnson, 2014, Fromkin et al., 2018]. Además, el Alfabeto Fonético Internacional (AFI) proporciona una notación alfabética para representar los sonidos fonéticos.

1.6. Fonología

La fonología se centra en el estudio de cómo los sonidos del habla forman patrones y construyen significado. Los fonemas son las unidades básicas de sonido que diferencian el significado de las palabras. Por ejemplo, en inglés, la "pɀ la "b"son fonemas distintos porque cambian el significado de las palabras en las que se encuentran. La fonología también examina las variaciones en la pronunciación de los sonidos en diferentes contextos y dialectos [Fromkin et al., 2018].

1.7. Morfología

La morfología se ocupa del estudio de la estructura interna de las palabras. Los morfemas son las unidades mínimas de significado que componen las palabras. Por ejemplo, en la palabra "deshacer", los morfemas son "des-", "hacerɀ er". La morfología también se interesa por los procesos de formación

de palabras, como la derivación, donde se agregan prefijos o sufijos a una palabra existente para formar una nueva palabra con un significado diferente [Johnson, 2014].

1.8. Sintaxis

La sintaxis es el estudio de cómo las palabras se combinan para formar frases y oraciones gramaticales. Examina las reglas y estructuras que determinan la organización de las palabras en una oración y cómo influyen en el significado. La sintaxis también se ocupa de la relación entre las palabras y las funciones que desempeñan dentro de una oración. Por ejemplo, en la oración .El perro persigue al gato", .el perro.es el sujeto, "persigue.es el verbo y .al gato.es el complemento directo [Johnson, 2014].

1.9. Semántica

La semántica es el estudio del significado de las palabras, las frases y las oraciones. Examina cómo se construye y se interpreta el significado en el contexto del lenguaje. La semántica también se interesa por los roles semánticos, que indican la función de cada entidad en una oración. Por ejemplo, en la oración .El niño cortó la cuerda con una navaja", .el niño.es el agente, "la cuerda.es el tema y üna navaja.es el instrumento [Johnson, 2014].

1.10. Pragmática

La pragmática se centra en cómo el contexto influye en la interpretación y el significado de las expresiones lingüísticas. Examina cómo se utilizan las expresiones lingüísticas en situaciones reales y cómo los hablantes interpretan el significado implícito. Por ejemplo, la oración "Hace frío aquí"puede interpretarse como una sugerencia implícita de cerrar las ventanas [Fromkin et al., 2018].

1.11. Procesamiento del Lenguaje Natural y Aprendizaje Automático

Aunque los seres humanos somos grandes usuarios del lenguaje, también somos muy malos para comprender y describir formalmente las reglas que rigen el lenguaje.

Entender y producir lenguaje utilizando computadoras es altamente desafiante. Los métodos más conocidos para lidiar con datos de lenguaje se basan en el aprendizaje automático supervisado.

El aprendizaje automático supervisado consiste en intentar inferir patrones y regularidades a partir de un conjunto de pares de entrada y salida preanotados (también conocido como conjunto de datos de entrenamiento).

1.12. Conjunto de Datos de Entrenamiento: Datos de NER CoNLL-2003

Cada línea contiene un token, una etiqueta de parte de la oración, una etiqueta de sintagma y una etiqueta de entidad nombrada.

```
U.N.          NNP  I-NP  I-ORG
official      NN   I-NP  O
Ekeus         NNP  I-NP  I-PER
heads         VBZ  I-VP  O
for           IN   I-PP  O
Baghdad       NNP  I-NP  I-LOC
.             .    O     O
```

[1]Fuente: `https://www.clips.uantwerpen.be/conll2003/ner/`

1.13. Desafíos del Lenguaje

Existen tres propiedades desafiantes del lenguaje: la discreción, la composicionalidad y la dispersión.

Discreción: no podemos inferir la relación entre dos palabras a partir de las letras que las componen (por ejemplo, hamburguesa y pizza).

Composicionalidad: el significado de una oración va más allá del significado individual de sus palabras.

Dispersión: la forma en que las palabras (símbolos discretos) pueden combinarse para formar significados es prácticamente infinita.

1.14. Example NLP Tasks

1.14.1. Topic Classification

Topic classification is an NLP task where a document is classified into one of several categories, such as sports, politics, gossip, or economy. Words in the documents provide strong hints about their topic. However, writing rules for this task is challenging due to the complexity of language. Data annotation, where readers categorize documents into topics, can help create training data for supervised machine learning algorithms. These algorithms learn patterns of word usage that aid in document categorization [Eisenstein, 2018].

1.14.2. Sentiment Analysis

Sentiment analysis is the application of NLP techniques to identify and extract subjective information from textual datasets. One common problem in sentiment analysis is message-level polarity classification (MPC), where sentences are automatically classified as positive, negative, or neutral. State-of-the-art solutions use supervised machine learning models trained on manually annotated examples [Mohammad et al., 2013].

In sentiment classification, supervised learning is often used, with Support Vector Machines (SVMs) being a popular choice. SVMs aim to find a hyperplane that separates the classes with the maximum margin, achieving the best separation between positive, negative, and neutral classes [Eisenstein, 2018].

Knowledge about linguistic structures is important for feature design and error analysis in NLP. Machine learning approaches in NLP rely on features that describe and generalize language use instances. Linguistic knowledge guides the selection and design of features, helping the machine learning algorithm find correlations between language use and the target labels [Bender, 2013].

NLP poses several challenges, including annotation costs, domain variations, and the need for continuous updates. Manual annotation is labor-intensive and time-consuming. Domain variations require different patterns to be learned for different corpora. Models trained on one domain may not work well on another. Additionally, NLP models can become outdated as language use evolves over time [Eisenstein, 2018].

En este capítulo, hemos explorado el desafío de entender y producir lenguaje utilizando computadoras. El aprendizaje automático supervisado es una de las principales técnicas utilizadas para abordar este desafío. Además, hemos discutido las propiedades desafiantes del lenguaje, como la discreción, la composicionalidad y la dispersión. Estos aspectos nos muestran la complejidad inherente al procesamiento del lenguaje natural y nos desafían a encontrar soluciones efectivas.

Bibliografía

[Bender, 2013] Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, 6(3):1–184.

[Eisenstein, 2018] Eisenstein, J. (2018). Natural language processing. Technical report, Georgia Tech.

[Fromkin et al., 2018] Fromkin, V., Rodman, R., and Hyams, N. (2018). *An introduction to language*. Cengage Learning.

[Goldberg, 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

[Johnson, 2014] Johnson, M. (2014). Introduction to computational linguistics and natural language processing (slides). 2014 Machine Learning Summer School.

[Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.