Sistema de Análisis de Texto No Estructurado

Julio Castillo, Marina Cardenas

Laboratorio de Investigación de Software LIS/Dpto. Ingeniería en Sistemas de Información/ Facultad Regional Córdoba/ Universidad Tecnológica Nacional

{ jotacastillo, ing.marinacardenas }@gmail.com

Resumen

En este proyecto se busca utilizar técnicas de aprendizaje automático (machine learning), especialmente utilizando Redes Neuronales Artificiales (RNA) para analizar texto (por ejemplo un artículo de diario) y en base a ello determinar la existencia de texto (oraciones o párrafos) que tengan el "mismo sentido" es decir que presenten la misma semántica, o bien oraciones/párrafos que estén semánticamente relacionadas entre sí. Este problema es comúnmente conocido como identificación y reconocimiento de parafrases. El fenómeno es particularmente difícil de detectar por procedimientos automáticos especialmente por la ambigüedad del lenguaje y por la gran variabilidad léxica que se utiliza para expresar las mismas ideas.

Palabras clave: análisis de texto, parafrases, machine learning.

Contexto

El presente proyecto se encuentra consolidado dentro de la línea de investigación relacionada al análisis de texto y es llevado a cabo en el Laboratorio de Investigación de Software LIS del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba.

El mismo, se enmarca dentro del Grupo GIA (Grupo de Inteligencia Artificial) de la UTN-FRC, el cual tiene como objetivo general el investigar técnicas, algoritmos de inteligencia artificial, entre los que se destacan el estudio de las redes neuronales, autómatas celulares, análisis y procesamiento de imágenes, minería de datos, y su aplicabilidad y resolución de problemas de las ciencias naturales y de las ciencias sociales. El grupo está integrado por doctores, ingenieros, licenciados, becarios, y pasantes.

De esta manera, se puede observar que se investigan técnicas de IA tanto desde el punto de vista teórico, como desde el punto de vista práctico.

Mediante este proyecto nos proponemos abordar el problema de reconocimiento de parafrases mediante técnicas aprendizaje automático por computadora (también machine learning), en especial las basadas en redes neuronales artificiales y máquinas de soporte vectorial.

Creemos que el desarrollo de nuevas técnicas de aprendizaje automático, junto con aplicación de técnicas existentes, permitirán realizar aportes en la tarea del reconocimiento de parafrases.

Objetivos Específicos

- Desarrollar un sistema que permita discernir entre una decisión de clasificación entre dos pares de textos, los cuales se podrán clasificar como: Hay existencia de parafrases (Yes), y No existencia de parafrases (No).

- Extender el sistema para que permita discernir entre una decisión de clasificación en tres vías: Hay existencia de parafrases (Yes), y se desconoce la existencia de parafrases en base a la información actual (Unknown), hay

contradicción (Contradiction), entre dos pares de textos.

• Aplicación de técnicas de machine learning: Se desarrollarán nuevas técnicas para la resolución del problema de identificación de parafrases basadas en clasificadores neuronales, y máquinas de soporte vectorial.

Introducción

Dos fragmentos de texto se consideran parafrases si son sintácticamente diferentes pero denotan la misma semántica.

De esta manera la reescritura del significado de una frase u oración utilizando otras palabras constituye una parafrase. Este término deriva del griego "forma de expresión adicional".

Nuestro estudio preliminar indica que los siguientes módulos formarían parte de un sistema de reconocimiento de parafrases básico:

1. Identificar fragmentos de texto (que en adelante llamaremos “textos”) con significado autónomo [3].

- Módulo de Alineación Léxica.
- Módulo de Adquisición de Parafrases.
- Módulo de Extracción de Características (formado por: características de alineación, características de dependencia, características de parafrases y semánticas) [6] [7].
- Módulo Clasificador (empleando redes neuronales, y máquinas kernel). [1]

2. Establecer relaciones de implicación entre textos, teniendo en cuenta que una implicación bidireccional entre dos fragmentos de textos diferentes es considerada una parafrase entre los mismos.

2.1. Aplicación de técnicas de machine learning.

La utilidad de un sistema de reconocimiento y generación de parafrases estriba en que sería posible utilizarlo como submódulo en muchas aplicaciones complejas de procesamiento y análisis de textos. Por ejemplo, podría utilizarse en un sistema de resumen automático (un tal sistema podría ayudarnos a eliminar párrafos de un texto que presentan la misma semántica eligiendo aquel de menor longitud, y así prescindir de información que es redundante, obteniendo como resultado un texto resumido), en traducción automática (una base de datos de parafrases ayudaría a mejorar la efectividad de las respuestas que producen los traductores automáticos como por ejemplo: Bing Translator o Google Translate), o bien podría utilizarse para detectar inconsistencias o información faltante o información adicional entre diversas paginas de Wikipedia que traten sobre un mismo tema u objeto, además estas técnicas pueden extrapolarse a idiomas diferentes, entre otras muchas aplicaciones.

Resultados Esperados

Como resultado del presente trabajo se pretende construir un sistema de reconocimiento y adquisición de parafrases que trabaje principalmente sobre texto en Inglés utilizando técnicas de aprendizaje automático.

La selección del idioma Inglés se debe principalmente a que la mayor parte de las herramientas y recursos como Ontologías, Tesauros, entre otros están disponibles para el idioma Inglés. Estos recursos son útiles en técnicas de procesamiento de texto y serán empleadas por nuestro sistema al momento de identificar, clasificar y generar parafrases de manera automática.

Formación de Recursos Humanos

Además de docentes, también participan de este proyecto, becarios y alumnos del último nivel la carrera de Ingeniería en Sistemas de Información de la UTN FRC,

próximos a recibirse y con perspectivas de iniciarse en una carrera de posgrado o doctorado, con lo cual, uno de los objetivos del proyecto es el contribuir a la formación de dichos alumnos.
El equipo de investigación y desarrollo de software, está formado por investigadores de la Universidad Tecnológica Nacional, Facultad Regional Córdoba, que a continuación se detallan:

- Actualmente el Lic.Ing. Julio Castillo está desarrollando su tesis de doctorado en Ciencias de la Computación en la Universidad Nacional de Córdoba en la temática propuesta en el presente proyecto, lo que esperamos que contribuya un aporte tanto a nivel académico como curricular en su formación de posgrado.

- Así mismo la Ing. Marina Cardenas está evaluando la posibilidad de desarrollar su tema de tesis de maestría y de especialidad (ambas en Ingeniería en Sistemas en la Universidad Tecnológica Nacional- FRC) en la misma temática con una variación del enfoque desde el punto de vista de los sistemas de Generación del Lenguaje Natural (NLG).

Adicionalmente en el proyecto participan alumnos becarios de la carrera Ingeniería en Sistemas de Información de la UTN-FRC.

Bibliografia

[1] Chris Brockett and William B. Dolan, Support Vector Machines for Paraphrase Identification and Corpus Construction, in Third International Workshop on Paraphrasing (IWP2005), Asia Federation of Natural Language Processing, 2005.

[2] Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment

[3]Chris Quirk, Chris Brockett, and William B. Dolan, Monolingual Machine Translation for Paraphrase Generation, Association for Computational Linguistics, July 2004.

[5] Judith Klavans and Philip Resnik.1996. The Balancing Act. Combining Symbolic and Statistical Approaches to Language. MIT Press.

[6]. D. Lin and P. Pantel. DIRT - Discovery of Inference Rules from Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages
323–328, 2001.

[7]. C. Monz and M. de Rijke. Light-Weight Entailment Checking for Computational Semantics. In P. Blackburn and M. Kohlhase, editors, Inference in Computational Semantics (ICoS-3), pages 59–72, 2001.

[8]. Xiaodong He, Jianfeng Gao, Chris Quirk, Patrick Nguyen, Arul Menezes, Robert Moore, Kristina Toutanova, Mei Yang, Bill dolan, Mu Li, Chi-Ho Li, Dongdong Zhang, Long Jiang, Ming Zhou, George Foster, Roland Kuhn, Jing Zheng, Wen Wang, Necip Fazil Ayan, Dimitra Vergyri, Nicolas Scheffer, and Andreas Stolcke, The MSR-NRC-SRI MT System for NIST Open Machine Translation 2008 Evaluation, in The 2008 NIST Open Machine Translation Evaluation Workshop, 2008.