Proposta recibida en outubro 2018 e aceptada para publicación en decembro 2018.

Explorando Métodos Non-Supervisados para Calcular a Similitude Semántica Textual

Exploring Unsupervised Methods to Sematic Textual Similarity

Pablo Gamallo
CiTIUS
Univ. de Santiago de Compostela
pablo.gamallo@usc.es

Martín Pereira-Fariña
Departamento de Filosofía e Antropoloxía
Universidade de Santiago de Compostela
martin.pereira@incipit.es

Resumo

Neste traballo preséntanse varios métodos non-supervisados para a detección da similitude semántica textual, os cales están baseados en modelos distribucionais e no parseado de dependencias. Os sistemas son avaliados mediante datasets empregados na ASSIN Shared Task, celebrada conxuntamente co PROPOR 2016. Os métodos máis básicos ofrecen un mellor comportamento que aqueles, mais complexos, que inclúen información sintáctico-semántica na análise das oracións. Por último, o uso de modelos distribucionais construidos automaticamente a partir de corpus ofrece resultados comparábeis ás estratexias que utilizan recursos léxicos externos construídos manualmente.

Palabras chave

similitude textual, análise de dependencias, extracción de información aberta

Abstract

This paper presents some unsupervised methods for detecting semantic textual similarity, which are based on distributional models and dependency parsing. The systems are evaluated using the dataset realased by the ASSIN Shared Task co-located with PROPOR 2016. The more basic methods offer better behavior than the more complex ones, which include syntactic-semantic information in sentence analysis. Finally, the use of distributional models built automatically from corpora provides results comparable to strategies that use external lexical resources built manually.

Keywords

textual similarity, dependency analysis, open information extraction

1 Introdución

As paráfrases defínense como pares de oracións que conteñen a mesma ou case a mesma información (Androutsopoulos & Malakasiotis, 2010). Polo tanto, o recoñecemento de paráfrases consiste no recoñecemento de oracións (ou pequenos fragmentos de texto) que teñen aproximadamente o mesmo significado nun contexto dado. Unha tarefa similar a á identificación de paráfrases é a Similitude Semántica Textual (SST), a cal busca determinar o grao de equivalencia semántica entre dous fragmentos de texto.

SST pode empregarse en moitas das tarefas do Procesamento da Linguaxe Natural (PLN), dende recuperación de información ata a detección automática de plaxio. Existen varios métodos de SST na bibliografía, que van dende métodos non-supervisados e con recursos lixeiros ata métodos supervisados e con recursos intensos.

O principal obxectivo deste traballo é describir e avaliar métodos non-supervisados de SST baseados en modelos distribucionais e aplicados ao portugués. Máis concretamente, compararemos estratexias non-supervisadas de recursos lixeiros con outras estratexias, tamén non-supervisadas, mais que utilizan recursos máis intensos, como tesaurus, redes de coñecemento, ou mesmo información sintáctica. Todos os experimentos son levados a cabo usando o datasets proporcionado por ASSIN Shared Task (*Avaliação de Similaridade Semântica e Inferência Textual*), celebrado conxuntamente con PROPOR 2016 (Fonseca et al., 2016).

Na seguinte sección (2), describimos os modelos de SST para o portugués. A continuación, na Sección 3 presentamos tres métodos non-supervisados diferentes. Na Sección 4 expoñemos e discutimos os resultados dos nosos experimentos; por último, na Sección 5, resumimos as nosas principais conclusións e propoñemos algunas ideas para o traballo futuro.

DOI: 10.21814/lm.10.2.275
This work is Licensed under a Creative Commons Attribution 4.0 License

2 Similitude semántica textual para o Portugués

SST é unha das dúas tarefas avaliadas no ASSIN (Fonseca et al., 2016). A outra subtarefa, inferencia textual, está fóra do ámbito deste traballo. A tarefa SST consiste en asignar un valor numérico (entre 1 e 5) a pares de oracións segundo o grao de similitude semántica entre elas: canto maior sexa o valor numérico, maior é o grao de similitude entre elas. Esta tarefa está inspirada pola *SemEval Task 2* sobre similitude semántica textual (Agirre et al., 2015, 2016). Na tarefa compartida sobre SST no *SemEval 2016*, enviáronse 119 sistemas diferentes, o que denota o enorme interese deste campo.

A inmensa maioría (todos menos un) dos sistemas presentados en ASSIN estaban baseados no uso de métodos supervisados. A mellor equipa (Hartmann, 2016) aplicou regresión lineal para adestrar un clasificador cuxas características son os valores da medida do coseno que representan o grao de similitude de cada para de oracións. Estas modélanse de dous xeitos diferentes: a adición de valores TF-IDF (cada palabra da oración é un valor TF-IDF) e a adición de vectores de valores distribucionais, onde cada palabra se representa como un vector contextual a aprendido mediante redes neuronais (Mikolov et al., 2013). As similitudes do coseno entre estes tipos de representacións son valores de entrada do clasificador básico.

O segundo mellor sistema (e o mellor para o dataset do portugués europeo, de Fialho et al. (2016), adestrou un clasificador baseado en modelos de regresión (*Kernel Ridge Regression*) cun número maior de rasgos que os outros sistemas, inlcuíndo distancias de edición entre cadeas de caracteres, o tamaño da maior subcadea común de caracteres, distintas métricas de similitude dependentes dos valores de ocorrencia de TF-IDF. En total, o sistema usou máis de 90 características.

A única estratexia non suspervisada no ASSIN é a chamada *Reciclagem* e foi proposta por Alves et al. (2016). Este sistema usa medidas de similitude baseadas nas relacións semánticas extraídas desde tesauros externos e recursos léxicos. O preprocesamento é realizado co etiquetador morfosintático de OpenNLP (Apache) e o lematizador LemPort (Rodrigues et al., 2014). Entre os recursos léxicos utilizados, destaca PAPEL (Oliveira et al., 2010), que consiste en relacións extraídas do dicionario *Porto Editora da Língua Portuguesa*, mediante a elaboración de regras baseadas en regularidades atopadas nas definicións do dicionario. Alén deste recurso, os experimentos realizados con Reciclagem inclúen outras redes de coñecimento con maior cobertura, nomeadamente, *CARTÃO* (Oliveira et al., 2011), que a súa vez consta doutros recursos como PAPEL e as relacións extraídas do *Dicionário Aberto* (Simões et al., 2012), así como diferentes variantes do WordNet portugués: *OpenWordNet.PT* (de Paiva et al., 2012) e *PULO* (Simões & Guinovart, 2014).

Neste traballo, avaliaremos varias estratexias non-supervisadas baseadas fundamentalmente en modelos distribucionais sobre o mesmo dataset empregado en ASSIN.

3 Similitude semántica textual non-supervisada

Nesta sección, definimos tres estratexias non-supervisadas: a máis básica baséase na semántica distribucional e na etiquetaxe morfo-sintáctica (*PoS tagging*), mentres que os outros métodos dependen da análise sintáctica alén de técnicas de extracción de información aberta (*Open Information Extraction).*

3.1 Similitude distribucional

Unha das estratexias máis simples e básicas para calcular a similitude entre dúas oracións consiste en sumar os valores de semellanza entre cada par de palabras que aparecen nas dúas oracións comparadas. Neste caso, só tomamos en conta palabras léxicas, é dicir, nomes, verbos e adxectivos. O valor de similitude calcúlase concretamente mediante a medida de coseno entre vectores que conforman matrices de palabras encapsuladas (*word embeddings*) pre-adestradas. O algoritmo é o seguinte: escollemos a oración máis curta e seleccionamos a primeira palabra léxica de dita oración. A seguir, calculamos a similitude do coseno entre a palabra escollida e todas as palabras léxicas que conforman a oración máis longa, sumando todos os valores de semellanza de maneira a obtermos a relevancia léxica da primeira palabra seleccionada con respeito á oración máis longa. Despois, realizamos a mesma operación para o resto de palabras da oración máis curta e calculamos a media dividindo a suma final polo número total de palabras léxicas que conforman a oración curta. Máis formalmente, dado o vector da palabra $\mathbf{p_s}$ pertencente a U_s, onde U_s é o conxunto de vectores de palabras léxicas da oración curta, a *relevancia léxica*, LR, de $\mathbf{p_s}$ dada a oración máis longa, calcúlase do seguinte xeito:

$$LR(\mathbf{p_s}, U_l) = \sum_{\mathbf{p_i} \in U_l}^{L} Cosine(\mathbf{p_s}, \mathbf{p_i}) \quad (1)$$

onde U_l é o conxunto dos vectores de palabras léxicas da oración máis longa e L é o número de palabras léxicas nesa mesma oración. Por conseguinte, o valor final de similitude (DSim) para o par U_s e U_l é a media de LR:

$$\text{DSim}(U_s, U_l) = \frac{\sum_{\mathbf{p_i} \in U_s}^{S} LR(\mathbf{p_s}, U_l)}{S} \quad (2)$$

onde S é o número de palabras léxicas na oración máis curta. Convén salientar que esta estratexia non codifica a información sobre a orde dos elementos da oración.

3.2 Extracción de proposicións básicas

DSim só toma en conta relacións semánticas ao nivel da palabra sen considerar fenómenos máis complexos como a orde das palabras ou as dependencias sintácticas entre elas. Para podermos tomar en conta estes fenómenos, desenvolvemos unha nova metodoloxía na que a estratexia de similitude definida previamente (DSim) se aplica a *proposicións básicas* extraídas das oracións, en vez de aplicarse directamente ás oracións. As proposicións básicas (ou tripletas) son relacións suxeito-verbo-obxecto identificadas e extraídas mediante técnicas de Extracción de Información Aberta (OIE) (Etzioni et al., 2011; Gamallo & García, 2015). Unha oración pode conter varias proposicións básicas, como por exemplo, a oración seguinte:

> *En maio de 2010, os partidos da oposición boicotaron as eleccións despois de acusacións de fraude electoral.*

Esta oración, despois dunha análise sintáctica en dependencias, pode dividirse en, polo menos, tres tripletas ou proposicións básicas, tal e como se mostra no cadro 1.

O método de cómputo de similitude baseada en proposicións, que chamamos BPROP, só toma en conta as palabras léxicas contidas nas proposicións extraídas. Deste xeito, pódese computar a similitude DSim comparando as tres proposicións do cadro 1 (extraídas da oración do noso exemplo) coa seguinte proposición (extraída da oración máis curta: *os partidos boicotaron as eleccións*):

subject	**relation**	**object**
partido	boicotar	eleccións

Os dous conxuntos de vectores de palabras léxicas elabóranse directamente das proposicións extraídas. A partir do exemplo citado, U_s (o conxunto de vectores de lemas da oración máis curta) é constituido mediante a selección dos lemas léxicos seguintes:

> {*partido, boicotar, elección*}

Pola outra banda, U_l (os vectores de lemas da oración máis longa), consta de:

> {*partido, oposición, boicotar, elección, maio, 2010, acusación, fraude, electoral*}

Estes conxuntos de vectores de lemas serven para calcular tanto a relevancia léxica (ecuación 1) como a similitude DSim (2). Polo tanto, a estratexia BPROP só determina que lemas están nos conxuntos comparados, mais non modifica o método de cómputo da similitude en si mesmo.

3.3 Estrutura de argumentos

A terceira estratexia que imos utilizar é moi similar a BPROP, mais en vez de extraer todas as posíbeis relacións suxeito-verbo-obxecto, o obxectivo da mesma é seleccionar a estrutura argumental principal de cada oración. Definimos a estrutura argumental principal dunha oración como aquela formada pola raíz (verbo principal) e os núcleos dos seus constituíntes directos. Deste xeito, a similitude baseada na estrutura argumental, que chamamos ARGSTR, calcúlase a partir das palabras léxicas que se encontran dentro do esqueleto estrutural extraído das oracións comparadas.

Como no caso de BPROP, a estratexia ARGSTR só modifica as listas de lemas que se van utilizar para computar a similitude DSim. No caso das dúas oracións comparadas no exemplo anterior, a lista correspondente á oración máis curta, U_s, sería a mesma que no caso anterior:

> {*partido, boicotar, elección*}

Pois tanto *partido* como *elección* son os núcleos dos constituíntes directos do verbo raíz: *boicotar*. No entanto, a lista da oración máis longa, U_l, é máis restritiva que na estratexia BPROP:

> {*partido, boicotar, elección, maio, acusación*}

O resto de lemas léxicos: *oposición*, *2010*, *fraude* e *electoral* non son constituíntes directos do verbo raíz senon doutros constituíntes da cláusula.

subject	relation	object
partido de oposición	boicotar	elección
partido de oposición	boicotar elección en	maio de 2010
partido de oposición	boicotar elección despois de	acusación de fraude electoral

Cadro 1: Tres proposicións básicas extraídas de: *En maio de 2010, os partidos da oposición boicotaron as eleccións despois de acusacións de fraude electoral*

4 Expermimentos

Para avaliar a calidade das estratexias definidas na sección previa no seu uso na captura de similitude semántica textual (SST), probámolas nos datasets fornecidos pola tarefa partillada ASSIN (Fonseca et al., 2016). Os experimentos foron realizados utilizando varios modelos semánticos pre-adestrados e disponíbeis publicamente, nomeadamente os modelos distribucionais transparantes e baseados en sintaxe descritos en Gamallo (2017).

O texto das oracións do test foi procesado con diferentes módulos de LinguaKit, unha suite lingüística mulilingüe e de código aberto (Gamallo & Garcia, 2017).[1] Máis concretamente, para podermos implementar as tres estratexias introducidas previamente, utilizamos o módulo de etiquetaxe morfo-sintáctica (Garcia & Gamallo, 2015), o parser de dependencias incluído en LinguaKit (Gamallo & Garcia, 2018), que son necesarios para desenvolver a estratexia ARGSTR, así como o módulo OIE de extracción de tripletas, (Gamallo & García, 2015), requirido por BPROP.

O cadro 2 mostra os valores, en termos de correlación de Pearson, obtidos polas tres estratexias non-supervisadas obxecto de estudo (DSim, ARGSTR, and BPROP), en base a tres listas de pares de oracións: portugués europeo, portugués brasileiro e a unión dos dous (Total). A cada par de oracións se lle asigna un valor entre 1 e 5, de xeito que canto maior é o valor maior é a semellanza entre as dúas oracións comparadas. Cada sistema é avaliado mediante a medición da correlación entre os valores anotados por persoas e os valores devolvidos polo sistema. O cadro tamén mostra na última fila os resultados atinxidos polo único sistema non-supervisado, *Reciclagem*, que participou na tarefa partillada ASSIN.

Como se pode comprobar, as estratexias máis básicas (DSIM e Reciclagem) son as que conseguen os mellores resultados. Ambas abordaxes utilizan información lingüística moi básica: lematización e recursos semánticos externos (modelos distribucionais pre-adestrados, no caso de DSim, e tesaurus externos, no casos de Reciclagem). Polo contrario, as dúas estratexias baseadas en información lingüística mais elaborada, nomeadamente análise sintáctica e extracción de información aberta (ARGSTR e BPROP) devolven valores moito máis baixos. Mesmo se vai ser preciso realizar unha análise de erros en profundidade, unha análise superficial dos mesmos lévanos a afirmar que os erros sintácticos provocados polo analizador son determinantes nos resultados finais destas dúas estratexias.

É preciso tamén salientar que hai unha importante diferenza entre DSim e Reciclagem. O primeiro utiliza modelos distribucionais automaticamente construídos a partir de corpus, mentres que o segundo utiliza relacións semánticas extraídas de recursos elaborados manualmente. En consecuencia, a estratexia inherente a DSim é completamente non-supervisada, mentres que o método utilizado por Reciclagem require unha supervisión distante pois, en última instancia, depende de tesaurus manuais.

5 Conclusións

Neste traballo probamos e avaliamos diferentes estratexias non-supervisadas para medir a similitude semántica textual. O estándar de referencia, cimentado unicamente no cálculo das palabras compartidas e similares, claramente mellora os resultados de métodos máis complexos enriquecidos con análise sintáctica e extracción básica de proposicións. Os resultados tamén mostran como o uso de modelos distribucionais dentro de estratexias non-supervisadas conseguen valores comparábeis aos que usan recursos léxicos externos construídos manualmente.

Para o traballo futuro, analizaremos con detalle os tipos de erros xerados nas técnicas baseadas na análise sintáctica co obxectivo de propor novas estratexias non-supervisadas para SST. Tamén avaliaremos estas técnicas con datasets orientados a outras tarefas máis alá da SST, tales como a identificación de paráfrases, as cales poderían ser máis sensíbeis á información sintáctica.

[1] https://github.com/citiususc/Linguakit

Sistemas	PT Europeo	PT Brasileiro	Total
DSim	**0.54**	0.56	0.53
ARGSTR	0.27	0.22	0.24
BPROP	0.29	0.24	0.26
Reciclagem	0.53	**0.59**	**0.54**

Cadro 2: Valores (correlación Pearson) devoltos polos nosos tres sistemas e máis pola estratexia non-supervisada (*Reciclagem*), que participou na tarefa compartida ASSIN.

Agradecementos

Este traballo foi financiado polo proxecto TelePares (MINECO, ref:FFI2014-51978-C2-1-R), e a Consellería de Cultura, Educación e Ordenación Universitaria (acreditación 2016-2019, ED431G/08 e Programa de Formación Posdoutoral da Xunta de Galicia 2016) e European Regional Development Fund (ERDF).

Referencias

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria & Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *9th International Workshop on Semantic Evaluation (SemEval)*, 252–263.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau & Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. En *10th International Workshop on Semantic Evaluation (SemEval)*, 497–511.

Alves, Ana Oliveira, Ricardo Rodrigues & Hugo Gonçalo Oliveira. 2016. ASAPP: alinhamento semântico automático de palavras aplicado ao português. *Linguamática* 8(2). 43–58.

Androutsopoulos, Ion & Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38. 135 – 187.

Apache. 2014. *Apache OpenNLP*. The Apache Software Foundation. http://opennlp.apache.org.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland & Mausam. 2011. Open Information Extraction: the Second Generation. En *International Joint Conference on Artificial Intelligence*, 3–10.

Fialho, Pedro, Ricardo Marques, Bruno Martins, Luísa Coheur & Paulo Quaresma. 2016. INESC-ID@ASSIN: Medição de similaridade semântica e reconhecimento de inferência textual. *Linguamática* 8(2). 33–42.

Fonseca, Erick Rocha, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Gamallo, Pablo. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation* 51(3). 727–743.

Gamallo, Pablo & Marcos García. 2015. Multilingual open information extraction. En *17th Portuguese Conference on Artificial Intelligence (EPIA)*, 711–722.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1).

Gamallo, Pablo & Marcos Garcia. 2018. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management* 54(6). 1244–1261.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. En *Languages, Applications and Technologies (CCIS)*, vol. 563, 65–75.

Hartmann, Nathan Siegle. 2016. Solo queue at ASSIN: combinando abordagens tradicionais e emergentes. *Linguamática* 8(2). 59–64.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. En *Conference of the North American Chapter of the ACL: Human Language Technologies*, 746–751.

Oliveira, Hugo Gonçalo, Leticia Antón Pérez, Hernani Pereira Costa & Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir

de dicionários electrónicos. *Linguamática* 3(2). 23–38.

Oliveira, Hugo Gonçalo, Diana Santos & Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e a sua avaliação. *Linguamática* 2(1). 77–93.

de Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian wordnet for reasoning. En *International Conference on Computational Linguistics (COLING)*, 353–360.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira & Paulo Gomes. 2014. LemPORT: a high-accuracy cross-platform lemmatizer for Portuguese. En *3rd Symposium on Languages, Applications and Technologies (SLATE)*, 267–274.

Simões, Alberto & Xavier Gómez Guinovart. 2014. Bootstrapping a Portuguese WordNet from Galician,Spanish and English wordnets. En *Second International Conference on Advances in Speech and Language Technologies for Iberian Languages (IberSPEECH)*, 239–248.

Simões, Alberto, Álvaro Iriarte Sanromán & José João Almeida. 2012. Dicionário-aberto: A source of resources for the portuguese language processing. En *Computational Processing of the Portuguese Language (PROPOR)*, 121–127.