Proposta recebida em Setembro 2016 e aceite para publicação em Novembro 2016.

INESC-ID@ASSIN: Medição de Similaridade Semântica e Reconhecimento de Inferência Textual

INESC-ID@ASSIN: Measuring Semantic Similarity and Recognizing Textual Entailment

Pedro Fialho
Universidade de Évora, INESC-ID
pedro.fialho@l2f.inesc-id.pt

Ricardo Marques
IST/UTL, INESC-ID
ricardo.sa.marques@tecnico.ulisboa.pt

Bruno Martins
IST/UTL, INESC-ID
bruno.g.martins@tecnico.ulisboa.pt

Luísa Coheur
IST/UTL, INESC-ID
luisa.coheur@l2f.inesc-id.pt

Paulo Quaresma
Universidade de Évora, INESC-ID
pq@di.uevora.pt

Resumo

Neste artigo apresentamos o sistema INESC-ID@ASSIN, o qual competiu no evento "Avaliação de Similaridade Semântica e Inferência Textual" (ASSIN) de 2016, nas tarefas de similaridade semântica e reconhecimento de paráfrases (i.e., inferência textual). O sistema INESC-ID@ASSIN aborda o problema de medir a similaridade entre frases como uma tarefa de regressão e aborda a inferência textual como uma tarefa de classificação. Embora o INESC-ID@ASSIN seja baseado essencialmente em características lexicais simples para deteção de paráfrases e reconhecimento de inferência textual, foram obtidos resultados promissores nesta avaliação conjunta.

Palavras chave

aprendizagem supervisionada, regressão, classificação

Abstract

In this article we present INESC-ID@ASSIN, a system that competed in the 2016 joint evaluation effort entitled *Avaliação de Similaridade Semântica e Inferência Textual* (ASSIN), in the tasks of semantic similarity and textual entailment recognition. INESC-ID@ASSIN addresses the problem of detecting sentence similarity as a regression task, and it addresses textual entailment as a classification task. Although INESC-ID@ASSIN relies mainly on simple lexical features for detecting paraphrases and recognizing textual entailment, promising results were achieved in this joint evaluation.

Keywords

supervised learning, regression, classification

1 Introdução

Detetar a quantidade e o tipo de similaridade entre duas frases é uma tarefa complexa de Compreensão de Língua Natural, principalmente devido à variabilidade lexical e sintática característica da língua natural. Detetar equivalência entre frases pode incluir a medição de semelhança semântica, e o problema está também relacionado com as tarefas de identificação de paráfrases ou de inferência textual.

A inferência textual pode ser definida como a tarefa de estimar a relação entre duas unidades de língua natural (por exemplo, entre duas frases), onde a veracidade de uma requer a veracidade da outra. Podemos dizer que de uma frase A se deduz a frase B se e somente se sempre que A é verdade B também é verdade.

Paráfrases são um tipo especial de inferência, nomeadamente inferência bidirecional. Uma paráfrase é uma espécie de equivalência semântica, responsável pela interligação de frases através da substituição de classes gramaticais e mantendo variáveis inalteradas entre as estruturas lexicais e sintáticas.

As tarefas de Identificação de Inferência Textual (RTE, do Inglês Recognizing Textual Entailment) e cálculo da similaridade semântica têm muitas aplicações práticas, sendo usadas em sistemas de pergunta-resposta, para extração de informação, sumarização ou tradução automática (MT, do Inglês Machine Translation), entre outros.

Neste artigo apresentamos o INESC-ID@ASSIN, um sistema que deteta paráfrases e faz inferência textual, baseado em aprendizagem automática supervisionada e que explora propriedades lexicais que relacionam duas frases. Detetar a quantidade de semelhança é conseguido com um modelo de regressão, enquanto o tipo de inferência é previsto com um classificador.

Avaliámos a nossa abordagem no contexto da ASSIN (Avaliação de Similaridade Semântica e Inferência Textual), uma tarefa de avaliação conjunta no PROPOR (Conferência Internacional sobre o Processamento Computacional do Português) de 2016. A tarefa ASSIN forneceu dados de treino e teste com exemplos em Português Europeu (PT-PT) e do Brasil (PT-BR).

O resto deste artigo está organizado da seguinte forma: A Secção 2 apresenta trabalhos relacionados. A Secção 3 apresenta o sistema INESC-ID@ASSIN e a Secção 4 detalha a avaliação e resultados. Finalmente, a Secção 5 conclui e indica trabalho futuro.

2 Trabalho relacionado

O aparecimento de tarefas conjuntas focadas no problema da RTE tem fomentado experiências com várias abordagens baseadas em dados/aprendizagem, aplicadas a tarefas semânticas (Dagan et al., 2009, 2013; Zhao et al., 2014; Bjerva et al., 2014). Particularmente, a disponibilidade de conjuntos de dados para aprendizagem supervisionada tornou possível formular o problema da RTE como uma tarefa de classificação, em que características são extraídas a partir dos exemplos de treino e utilizadas pelos algoritmos de aprendizagem automática na construção de um classificador, que é finalmente aplicado aos dados de teste.

Abordagens recentes para RTE ou para a identificação de paráfrases utilizam algoritmos de aprendizagem automática (por exemplo, classificadores lineares) com uma variedade de características, baseadas em comparações sobre padrões lexicais, sintáticos e/ou semânticos, contagem de co-ocorrências em documentos, e regras de primeira ordem para reescrita sintática.

Diferentes abordagens têm sido formuladas, muitas vezes envolvendo a combinação de características como as acima descritas. Uma abordagem simples é a estratégia saco-de-palavras, em que a semelhança de um par de frases é calculada utilizando a similaridade do cosseno entre representações vetoriais. Se o valor da similaridade é superior a um limiar pré definido (estabelecido manualmente ou aprendido através de dados) as frases são classificadas como paráfrases.

Zhang & Patrick (2005) propuseram um método de classificação em que o par de frases é simplificado para formas canónicas através de regras para alterar a voz passiva para ativa, entre outras. Utilizando árvores de decisão, os autores exploram características baseadas em comparações lexicais, tais como a distância de edição entre símbolos (e.g., letras ou palavras).

Além de utilizar características de comparação lexical, autores como Kozareva & Montoyo (2006) ou Ul-Qayyum & Wasif (2012) propuseram abordagens baseadas em classificação utilizando uma combinação de características lexicais, semânticas e heurísticas (por exemplo, padrões de negação) para auxiliar a deteção de falsas paráfrases.

Os métodos utilizados na maioria das anteriores abordagens funcionam ao nível das frases, mas visto que as paráfrases utilizam tipicamente sinónimos ou outras formas de palavras relacionadas, autores como Mihalcea et al. (2006) ou Fernando & Stevenson (2008) desenvolveram métodos de similaridade ao nível de palavras para determinar se uma frase é paráfrase de outra. Estes métodos são baseados em medidas de similaridade palavra-a-palavra (por exemplo, métricas baseadas em dados que utilizem a WordNet). Métodos baseados em alinhamentos (como os formulados para sumarização ou tradução) são também usuais.

Madnani et al. (2012) propuseram uma abordagem baseada em métricas para alinhamento de sequências de caracteres, utilizadas em tradução automática (MT). Embora o uso de métricas de MT para a tarefa de identificação de paráfrases não seja novidade (Finch et al., 2005), o mérito dos autores está na re-avaliação dessas métricas, conjuntamente com a criação de novas métricas, alcançando um dos melhores resultados sobre o conhecido Microsoft Research Paraphrase Corpus (Dolan et al., 2004).

Pakray et al. (2011) descrevem uma abordagem lexical e sintática para resolver o problema da RTE. Este método resulta da composição de vários módulos, nomeadamente módulos de pré-processamento, similaridade lexical e similaridade sintática.

Tsuchida & Ishikawa (2011) propuseram um sistema RTE que usa métodos de aprendizagem automática com características baseadas em informação lexical e ao nível das estruturas predicado-argumento. A ideia subjacente é delimitar os pares texto-hipótese identificados como tendo inferência textual, mas que na verdade não

têm, ou seja, falsos positivos classificados pelo módulo de nível lexical podem ser rejeitados pelo módulo de nível da frase.

É importante notar que os trabalhos anteriores normalmente correspondem a métodos que são independentes do idioma pelo uso de estratégias simples, tal como a contagem *n*-gramas. Da maioria das abordagens RTE descritas também se conclui que os módulos lexicais alcançam melhores resultados do que os módulos sintáticos e baseados na estrutura de frases.

As mais recentes abordagens a estes problemas dependem de recursos dependentes do idioma e, como seria de esperar, focam-se na língua Inglesa, explorando modelos de semântica distribuída, utilizando recursos como word embeddings (Cheng & Kartsaklis, 2015). Apenas muito recentemente foram publicados recursos que permitiriam replicar algumas destas experiências tendo em conta o Português (por exemplo, (Rodrigues et al., 2016)).

3 INESC-ID@ASSIN

Os modelos de regressão/classificação gerados no contexto do INESC-ID@ASSIN foram baseados no formalismo dos *kernel methods* e usam várias métricas de similaridade. Vários estudos anteriores, na área de Processamento de Língua Natural (NLP, do Inglês Natural Language Processing) e também em outros domínios, usaram métodos semelhantes para combinar múltiplas métricas de similaridade no contexto de obter a semelhança entre objetos (Martins, 2011; Madnani et al., 2012).

As métricas usadas para extrair características dos dados têm em conta, em especial, contribuições da informação lexical. Algumas destas métricas inspiram-se em estudos focados na identificação de paráfrases; outras em estudos relativos a RTE. Várias formas de representação do texto são tidas em conta (minúsculas, Metaphone, etc.).

Os recursos utilizados no INESC-ID@ASSIN são explicados nas seguintes secções e descritos mais detalhadamente em Marques (2015). Uma máquina de suporte de vectores (do Inglês *Support Vector Machine* (SVM)) foi utilizada para a classificação (RTE e identificação de paráfrases) e um modelo do tipo *Kernel Ridge Regression* (KRR) foi utilizado para obter valores contínuos (quantificação de similaridade). Usamos as implementações SVM/KRR do pacote de ferramentas scikit-learn[1], para Python.

[1] http://scikit-learn.org/

3.1 Similaridade lexical

As características de comparação lexical consideradas no INESC-ID@ASSIN são as seguintes:

1. **Maior Subsequência Comum**. O tamanho da maior subsequência comum (LCS) entre o texto e a hipótese. O valor é fixado entre 0 e 1, dividindo o tamanho da LCS pelo tamanho da frase mais longa.
2. **Distância de edição**. A distância mínima de edição/alteração entre símbolos (letras ou palavras) do texto e da hipótese.
3. **Comprimento**. A diferença absoluta de comprimento (número de símbolos) entre o texto e a hipótese. Os comprimentos máximo e mínimo são também considerados (separadamente) como características.
4. **Similaridade por Cosseno**. A similaridade do cosseno entre o texto e a hipótese, com base no número de ocorrências de cada palavra no texto/hipótese (a representação usa a frequência dos termos nos vetores associados a cada documento). A fórmula do cosseno é mostrada na Equação 1.

$$\cos(s_1, s_2) = \frac{\overrightarrow{V}(s_1) \cdot \overrightarrow{V}(s_2)}{||\overrightarrow{V}(s_1)|| \times ||\overrightarrow{V}(s_2)||} \quad (1)$$

   O resultado é um número contínuo entre 0 e 1. Quanto maior o valor, maior a semelhança no par texto-hipótese.
5. **Similaridade de Jaccard**. A similaridade de Jaccard entre o texto e a hipótese. O valor retornado é um número contínuo entre 0 e 1, onde 1 significa que as frases são iguais, e 0 que são totalmente diferentes. O coeficiente de similaridade de Jaccard é usado para comparar a semelhança e diversidade de conjuntos. Mede a semelhança entre conjuntos finitos, e é definido como a divisão entre o número de elementos na intersecção e na união dos conjuntos. A similaridade de Jaccard entre dois conjuntos de palavras s_1 e s_2 é assim definida da seguinte forma:

$$\text{Jaccard}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \quad (2)$$

6. **Soft TF-IDF**. A métrica Soft TF-IDF mede a similaridade entre representações vetoriais das frases, mas considerando uma métrica de similaridade interna para encontrar palavras equivalentes. A métrica Jaro-Winkler para

similaridade entre palavras, com um limiar de 0.9, é utilizada como métrica de similaridade interna. A distância $\text{Jaro}(s_1, s_2)$ entre duas sequências s_1 e s_2 é:

$$\text{Jaro}\,(s_1, s_2) = \begin{cases} 0 & \text{se } m = 0 \\ \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{se } m \neq 0 \end{cases} \quad (3)$$

Na equação, m é o número de caracteres coincidentes e t é metade do número de transposições. A métrica Jaro-Winkler modifica a métrica Jaro adicionando-lhe mais peso quando há um prefixo em comum. Este melhoramento define 2 variáveis: (1) *PL*, o comprimento do maior prefixo comum entre duas sequências, com um limite de 4, e (2) *PW*, o peso a dar ao prefixo.

$$\text{JaroWinkler}\,(s_1, s_2) = (1 - \text{PL} \times \text{PW}) \times \text{Jaro}\,(s_1, s_2) + \text{PL} \times \text{PW} \quad (4)$$

3.2 Características sobre RTE

As características inspiradas em estudos com o foco em RTE são as seguintes:

1. **Sobreposição NE**. A similaridade de Jaccard considerando apenas entidades mencionadas (NE – do Inglês Named Entities). Para simplificar, entidades mencionadas são todas as palavras que contêm letras maiúsculas.
2. **Sobreposição NEG**. A similaridade de Jaccard considerando apenas palavras negativas. As palavras negativas são: *não*, *nunca*, *jamais*, *nada*, *nenhum*, *ninguém*.
3. **Sobreposição Modal**. A similaridade de Jaccard considerando apenas palavras modais. As palavras modais são: *podia*, *poderia*, *dever*, *deve*, *devia*, *deverá*, *deveria*, *faria*, *possível*, *possibilidade*, *possa*.

3.3 Características sobre paráfrases

As características inspiradas em estudos sobre identificação de paráfrases são as seguintes:

1. **BLEU**. Esta métrica de MT corresponde à quantidade de sobreposições em n-gramas, para diferentes valores de n, entre duas frases, ajustada por uma penalização relativa ao seu comprimento (Papineni et al., 2002). O maior n que utilizámos foi 3, para a cobertura de frases curtas, visto que é sugerido em Papineni et al. (2002) que este valor produz um desempenho semelhante, em comparação com o valor clássico de 4-gramas (BLEU-4).
2. **METEOR**. Esta métrica é uma variação do BLEU com base na média harmónica da precisão e cobertura de unigramas, tendo a cobertura maior peso do que a precisão (Banerjee & Lavie, 2007).
3. **TER**. A Taxa de Erros de Tradução (TER) é uma extensão da Taxa de Erros em Palavras (ou Word Error Rate — WER), que é uma métrica simples baseada em programação dinâmica e que é definida como o número de alterações necessárias para transformar uma sequência noutra. A TER inclui um algoritmo heurístico para lidar com transposições, além de inserções, remoções e substituições (Snover et al., 2006).
4. **NCD**. A Distância de Compressão Normalizada (NCD) é uma forma geral de medir a similaridade entre dois objetos (Li et al., 2004). A ideia subjacente é que ao compactar duas sequências s_1 e s_2 somente a informação sobreposta é extraída.
5. **ROUGE-N**. Sobreposição de n-gramas com base em estatísticas de co-ocorrência (Lin & Hovy, 2003).
6. **ROUGE-L**. Uma variação da métrica ROUGE-N com base no comprimento da maior subsequência de palavras comum (Lin & Och, 2004).
7. **ROUGE-S**. Uma variação da métrica ROUGE-N baseada em skip-bigrams (ou seja, bigramas/pares de palavras, pela ordem em que ocorrem na frase, e possibilitando intervalos entre as palavras) (Lin & Och, 2004).

3.4 Características numéricas

A inspiração para estas características numéricas é simples: frases que se referem às mesmas entidades, mas com números diferentes, são suscetíveis de ser contraditórias. O cálculo desta característica é simples, resultando da multiplicação de 2 similaridades de Jaccard. Uma entre os caracteres numéricos no par texto-hipótese, e outra entre as palavras em torno de tais caracteres numéricos. O resultado é um valor contínuo entre 0 e 1, onde 0 indica que as frases são possivelmente contraditórias.

3.5 Representações de texto

As características anteriormente descritas são aplicadas a diferentes representações das frases. Nomeadamente, considerámos as seguintes representações:

1. **Símbolos originais**.
2. **Símbolos em minúsculas**.
3. **Símbolos em minúsculas sem variações terminais (obtidos pela aplicação de um algoritmo de *stemming*)**.
4. **Agrupamentos de palavras**. O algoritmo de Brown para o agrupamento de palavras é um método aglomerativo que agrega palavras numa árvore binária de classes (Turian et al., 2010), através de um critério baseado na probabilidade logarítmica de um texto perante um modelo de língua baseado em classes.

   O procedimento de agrupamento de Brown foi aplicado a uma coleção de documentos noticiosos do jornal Português *Público*, do qual resultaram 1001 agrupamentos. Nesta representação, as palavras/símbolos são substituídos pelas classes correspondentes.
5. **Double Metaphone**. Foi utilizado um algoritmo bem conhecido para codificar palavras com base na sua fonética, interpretando cada palavra como uma combinação dos sons de 12 consoantes. No entanto, importa referir que o algoritmo Double Metaphone (Philips, 1990) é baseado na pronúncia Inglesa, sendo mais adequado para codificar palavras em inglês e palavras estrangeiras tipicamente utilizadas nos Estados Unidos.
6. **Trigramas de caracteres**. Os trigramas são um caso especial do conceito de n-grama, onde n é 3. Os trigramas de caracteres são usados como termos-chave numa representação da frase, à semelhança de como as palavras são usadas como termos-chave para representar um documento.

Os nossos modelos combinam características com base nestas diferentes representações, considerando um total de 96 características. Algumas características não são adequados para serem combinadas com algumas representações, tal como a característica numérica com a representação Double Metaphone. As combinações consideradas são descritas na Tabela 1.

Feature	O	L	S	C	DM	T
LCS	✓	✓	✓	✓	✓	
D. de edição	✓	✓	✓	✓	✓	
Cosseno	✓	✓	✓	✓	✓	✓
C. Absoluto	✓	✓	✓	✓	✓	
C. Máximo	✓	✓	✓	✓	✓	
C. Mínimo	✓	✓	✓	✓	✓	
Jaccard	✓	✓	✓	✓	✓	✓
Soft TF-IDF	✓	✓	✓			
NE	✓	✓	✓	✓	✓	✓
NEG	✓	✓	✓	✓	✓	✓
Modal	✓	✓	✓	✓	✓	✓
BLEU-3	✓	✓	✓	✓	✓	
METEOR	✓	✓	✓	✓	✓	
ROUGE N	✓	✓	✓	✓	✓	
ROUGE L	✓	✓	✓	✓	✓	
ROUGE S	✓	✓	✓	✓	✓	
TER	✓	✓	✓	✓	✓	
NCD	✓	✓	✓	✓	✓	
Numérica	✓	✓	✓			

Tabela 1: Combinação de características com representações, onde O, L, S, C, DM e T correspondem a símbolos originais, minúsculas, sem terminações, agrupamentos, Double Metaphone e trigramas, respetivamente.

4 Avaliação

O INESC-ID@ASSIN foi avaliado no conjunto de dados ASSIN para medir o seu desempenho na tarefa de quantificar automaticamente a similaridade semântica e tipo de inferência textual.

Reportamos resultados de 2 configurações distintas, uma utilizando um kernel polinomial em modelos SVM e KRR e outra utilizando um kernel linear. Para os modelos lineares, as características mais informativas também são reportadas.

Cada experiência gerou resultados para 3 configurações diferentes, em ambas as tarefas e para dados de teste portugueses e brasileiros.

Além disso, também medimos o desempenho ao treinar o nosso sistema com uma variedade do Português e testar com a outra.

As configurações diferem nos dados utilizados para treino dos algoritmos de aprendizagem. Um desses conjuntos de dados corresponde à expansão do ASSIN com frases traduzidas automaticamente desde o corpus SICK (Marelli et al., 2014), enquanto que as restantes configurações usam partições do ASSIN original.

4.1 Descrição da Tarefa

O ASSIN contém 10000 pares de frases recolhidas de Google News, particionados em conjuntos de treino e teste, com um número de exemplos portugueses e brasileiros igualmente distribuído por cada conjunto. Cada par de frases é anotado para similaridade semântica e inferência textual.

A similaridade semântica é um valor contínuo de 1 a 5, de acordo com as seguintes diretrizes sobre as frases de um par:

1. Completamente diferentes, sobre diferentes assuntos;
2. Não relacionadas, mas mais ou menos sobre o mesmo assunto;
3. Algo relacionadas. Podem descrever factos diferentes, mas partilham alguns detalhes;
4. Fortemente relacionadas, mas alguns detalhes são diferentes;
5. Essencialmente a mesma coisa.

A anotação da inferência textual é uma atribuição categórica usando classes que identificam inferência, paráfrase ou nenhuma relação.

O ASSIN define 2 tarefas para quantificar/calcular a similaridade semântica e classificar o tipo de inferência textual. O desempenho é medido separadamente para as variantes de Portugal e do Brasil.

4.2 Treinar com mais dados

Experimentámos utilizar métodos de MT para expandir o conjunto de dados ASSIN original com novas frases de um conjunto de dados em Inglês, visto que mais dados normalmente conduzem a melhores resultados.

O conjunto de dados SICK (Marelli et al., 2014) é muito semelhante ao ASSIN, em tamanho e tipo de anotações. No entanto, é baseado em legendas de imagens e vídeos, obtidas por crowdsourcing, logo representa menor variabilidade linguística mas mais similaridade entre pares (ou seja, mais pares similares).

O SICK foi traduzido para Português, usando um programa Python assente no serviço de tradução online Microsoft Bing, e conjugado com os conjuntos de treino em português europeu e brasileiro. Assim, adicionamos 9191 exemplos do SICK aos 6000 exemplos do ASSIN, para uma das configurações.

4.3 Resultados

A nossa abordagem à tarefa ASSIN foi avaliada utilizando o coeficiente de Pearson e o erro quadrático médio (MSE) como métricas para similaridade semântica, e com a Exatidão e a medida F1 para RTE.

Consideramos 3 configurações/tentativas diferentes para a nossa abordagem, que diferem na quantidade de dados de treino que são usados, nomeadamente:

1. PT-PT or PT-BR: treinar apenas com dados da mesma variedade de Português (Europeu ou do Brasil, respetivamente) dos dados de teste (3000 exemplos).
2. AllPT: juntar os dados de ambas as variedades para treino, independentemente do teste pretendido (6000 exemplos).
3. PT+BingSICK: usar ambas as variedades e os dados do SICK traduzido para treino (15191 exemplos, dos quais 9191 são do SICK).

Estas configurações foram avaliadas nos dados de teste europeus e brasileiros, embora na entrega oficial só tenha sido avaliado o teste europeu. Na entrega oficial, PT com um kernel polinomial foi a nossa melhor configuração (nos dados de teste europeus). No entanto, devido a um problema no software (agora resolvido) os valores oficiais foram inferiores aos apresentados na Tabela 2.

Os resultados para a nossa abordagem à tarefa ASSIN, recorrendo a um kernel polinomial, são apresentados nas Tabelas 2 e 3.

	Similaridade		**RTE**	
Treino	**Pearson**	**MSE**	**Exatidão**	**F1**
PT-PT	0.74	0.60	83.55%	0.68
AllPT	0.74	0.60	83.95%	0.69
PT+BingSICK	0.72	0.68	80.70%	0.59

Tabela 2: Resultados da avaliação, com um kernel polinomial e considerando todas as características — teste europeu.

	Similaridade		**RTE**	
Treino	**Pearson**	**MSE**	**Exatidão**	**F1**
PT-BR	0.73	0.36	85.45%	0.64
AllPT	0.73	0.36	85.70%	0.66
PT+BingSICK	0.70	0.40	84.30%	0.58

Tabela 3: Resultados da avaliação, com um kernel polinomial e considerando todas as características — teste brasileiro.

Os resultados para a nossa abordagem à tarefa ASSIN, recorrendo a um kernel linear, são apresentados nas Tabelas 4 and 5.

	Similaridade		RTE	
Treino	**Pearson**	**MSE**	**Exatidão**	**F1**
PT-PT	0.73	0.62	84.90%	0.71
AllPT	0.74	0.61	84.05%	0.68
PT+BingSICK	0.70	0.73	77.10%	0.47

Tabela 4: Resultados da avaliação, com um kernel linear e considerando todas as características — teste europeu.

	Similaridade		RTE	
Treino	**Pearson**	**MSE**	**Exatidão**	**F1**
PT-BR	0.73	0.36	85.35%	0.55
PT	0.73	0.36	85.85%	0.66
PT+BingSICK	0.70	0.42	82.60%	0.46

Tabela 5: Resultados da avaliação, com um kernel linear e considerando todas as características — teste brasileiro.

O desempenho com um kernel linear é semelhante ao de um kernel polinomial, mas a vantagem da maior dimensionalidade do espaço de um kernel polinomial é realçada quando existem mais dados, como pode ser visto na queda de desempenho dos modelos lineares quando se utiliza o conjunto de dados expandido com MT (em particular no MSE e F1), comparando com os resultados obtidos com um kernel polinomial.

Destes resultados podemos concluir que utilizar dados de treino selecionados/verificados (manualmente) pode melhorar ligeiramente o desempenho, enquanto que dados de treino não filtrados (repetitivos e com erros lexicais ou sintáticos resultantes de MT) prejudica o desempenho da nossa abordagem.

Comparando os resultados por tabela, a configuração que mais consistentemente tem os melhores resultados é a AllPT, tanto para RTE como para medição da similaridade. Considerando todas as tabelas, o nosso sistema tem melhor desempenho nos dados da variante do Brasil.

Os restantes sistemas que participaram na tarefa ASSIN obtiveram resultados inferiores aos apresentados. Barbosa et al. (2016) experimenta SVM e redes neuronais em características baseadas em word embeddings, e apresenta uma visão geral dos resultados obtidos por todos os sistemas que participaram no ASSIN.

Em (Hartmann, 2016) são utilizadas características baseadas em conjuntos de palavras (logo esparsas), onde também figuram os word embeddings. Este sistema obteve os resultados mais próximos dos descritos neste artigo, embora só tenha participado na medição de similaridade semântica.

A abordagem de Freire et al. (2016) introduz um conjunto de ferramentas para sistemas de similaridade entre frases, instanciado com semântica distribuída e conhecimento da WordNet. Este sistema também não participou na medição de similaridade semântica.

Por último, o sistema de Alves et al. (2016) apresenta uma abordagem não supervisionada, individualmente e como característica de uma abordagem supervisionada. Os piores resultados são da abordagem não supervisionada, enquanto que a supervisionada atingiu resultados semelhantes aos de Barbosa et al. (2016), e os mais próximos dos resultados reportados neste artigo relativamente a RTE.

Experimentámos também compreender o desempenho dos modelos treinados com uma variedade de Português e testados com a outra variedade. Como apresentado na Tabela 6, compreender uma variedade do Português conhecendo apenas a outra é melhor do que utilizando o conjunto de dados SICK, traduzido automaticamente pelo sistema Bing. Para simplificar, só é apresentada a experiência com kernels polinomiais, mas com kernels lineares foram obtidos resultados semelhantes.

	Similaridade		RTE	
Treino	**Pearson**	**MSE**	**Exatidão**	**F1**
PT-BR	0.73	0.63	82.70%	0.64
PT-PT	0.72	0.37	84.30%	0.66

Tabela 6: Variando o conjunto de treino e testando com a outra/restante variedade do Português, utilizando um kernel polinomial e todas as características.

4.4 Melhores características

Utilizamos o método Recursive Feature Elimination, tal como implementado no scikit-learn, para obter as 10 melhores características com a configuração PT (i.e., a que produziu os melhores resultados), para cada tarefa (RTE e quantificação de similaridade).

Este é um método para seleção de características com base no seu peso relativamente ao modelo. Como o scikit-learn só representa os pesos das característica em modelos com kernels lineares, apenas aplicamos seleção de características nos nossos modelos lineares.

As 10 melhores características para RTE (classificação) são:

- Soft TF-IDF, em símbolos originais;
- Jaccard, sobre Double Metaphone;
- Jaccard, sobre símbolos em minúsculas sem variações terminais;
- Comprimento Absoluto, em Double Metaphone;
- LCS, sobre símbolos em minúsculas sem variações terminais;
- Numérica, em símbolos originais;
- Sobreposição NE, em Double Metaphone;
- ROUGE-N, em símbolos originais;
- ROUGE-L, sobre símbolos em minúsculas sem variações terminais;
- TER, sobre símbolos em minúsculas sem variações terminais.

As 10 melhores características para quantificação de similaridade (regressão) são:

- Similaridade do Cosseno, em símbolos originais;
- Soft TF-IDF, em símbolos originais;
- Jaccard, em Double Metaphone;
- Jaccard, sobre símbolos em minúsculas sem variações terminais;
- Jaccard, em trigramas de caracteres;
- Numérica, sobre símbolos em minúsculas sem variações terminais;
- Sobreposição NE, em Double Metaphone;
- ROUGE-N, sobre símbolos originais;
- ROUGE-N, em agrupamentos de palavras;
- ROUGE-S, sobre símbolos em minúsculas sem variações terminais.

As características baseadas em similaridade lexical contribuem para os melhores resultados de ambas as tarefas, em especial se se tiver em conta as representações que mantêm os símbolos da frase, como comprovado pela predominância destas métricas e representações entre as 10 melhores características. A única característica baseada em RTE que teve um desempenho relevante é a Sobreposição NE, sobre a representação de texto processado pelo algoritmo Double Metaphone.

5 Conclusões e trabalho futuro

Este trabalho tem por foco as tarefas de RTE e de quantificação de similaridade textual, abordando as mesmas através da aplicação de várias características baseadas em trabalhos anteriores para RTE e identificação de paráfrases - essencialmente métricas provenientes dos domínios de MT e sumarização. Estas características, juntamente com outras relativas a similaridade entre sequências e aspetos numéricos, representam uma nova abordagem que se afasta da mais recente tendência da área, que essencialmente se foca em sistemas baseados em alinhamentos semânticos e correspondência entre relações binárias.

Como trabalho futuro, iremos começar por comparar o desempenho do sistema INESC-ID@ASSIN com variantes, usando os mesmos algoritmos de aprendizagem, aplicados a características mais complexas baseadas em representações sintáticas/semânticas e baseadas em fontes de conhecimento enriquecidas.

Agradecimentos

Este trabalho foi suportado por fundos nacionais através da Fundação para a Ciência e a Tecnologia (FCT), através do projeto com referência UID/CEC/50021/2013. O trabalho foi ainda suportado pelo projeto internacional RAGE com referência H2020-ICT-2014-1/644187 e pelo projeto LAW-TRAIN com referência H2020-EU.3.7.-653587.

Referências

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

Banerjee, Satanjeev & Alon Lavie. 2007. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. Em *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 228–231.

Barbosa, Luciano, Paulo Cavalin, Victor Guimarães & Matthias Kormaksson. 2016. Blue Man Group no ASSIN: Usando representações distribuídas para similaridade semântica e inferência textual. *Linguamática* 8(2). 15–22.

Bjerva, Johannes, Johan Bos, Rob van der Goot & Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity.

Em *Proceedings of the International Workshop on Semantic Evaluation*, 642–646.

Cheng, Jianpeng & Dimitri Kartsaklis. 2015. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. Em *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1531–1542.

Dagan, Ido, Bill Dolan, Bernardo Magnini & Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering* 15(04). i–xvii.

Dagan, Ido, Dan Roth, Mark Sammons & Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies* 6(4). 1–220.

Dolan, Bill, Chris Quirk & Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. Em *Proceedings of the International Conference on Computational Linguistics*, s. pp.

Fernando, Samuel & Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. Em *Proceedings of the Annual Research Colloquium on Computational Linguistics in the UK*, s. pp.

Finch, Andrew, Young-Sook Hwang & Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. Em *Proceedings of the International Workshop on Paraphrasing*, 17–24.

Freire, Jânio, Vládia Pinheiro & David Feitosa. 2016. FlexSTS: Um framework para similaridade semântica textual. *Linguamática* 8(2). 23–31.

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

Kozareva, Zornitsa & Andres Montoyo. 2006. Paraphrase identification on the basis of supervised machine learning techniques. Em *Proceedings of the International Conference on Advances in Natural Language Processing*, 524–533.

Li, Ming, Xin Chen, Xin Li, Bin Ma & Paul Vitányi. 2004. The similarity metric. *Information Theory, IEEE Transactions on* 50(12).

Lin, Chin-Yew & Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. Em *Proceedings of the Conference of the North American Chapter of the ACL on Human Language Technology*, 71–78.

Lin, Chin-Yew & Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. Em *Proceedings of the Annual Meeting of ACL*, s. pp.

Madnani, Nitin, Joel Tetreault & Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. Em *Proceedings of the Conference of the North American Chapter of ACL*, 182–190.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi & Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. Em *Proceedings of the International Conference on Language Resources and Evaluation*, 216–223.

Marques, Ricardo. 2015. *Detecting contradictions in news quotations*: IST, University of Lisbon. Tese de Mestrado.

Martins, Bruno. 2011. A supervised machine learning approach for duplicate detection over gazetteer records. Em *Proceedings of the International Conference on GeoSpatial Semantics*, 34–51.

Mihalcea, Rada, Courtney Corley & Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. Em *Proceedings of the National Conference on Artificial Intelligence*, 775–780.

Pakray, Partha, Sivaji Bandyopadhyay & Alexander Gelbukh. 2011. Textual entailment using lexical and syntatic similarity. *Internacional Journal of Artificial Intelligence and Applications* 2(1). 43–58.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. Em *Proceedings of the Annual Meeting of ACL*, 311–318.

Philips, L. 1990. Hanging on the metaphone. *Computer Language Magazine* 7(12). 39–44.

Rodrigues, João António, António Branco, Steven Neale & João Ricardo Silva. 2016. Lx-dsemvectors: Distributional semantics models for portuguese. Em *Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016*, 259–270.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. Em *Proceedings of the Conference of the Association for Machine Translation in the Americas*, 223–231.

Tsuchida, Masaaki & Kai Ishikawa. 2011. A method for recognizing textual entailment using lexical-level and sentence structure-level features. Em *Proceedings of the Text Analysis Conference*, s. pp.

Turian, Joseph, Lev Ratinov & Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. Em *Proceedings of the Annual Meeting of ACL*, 384–394.

Ul-Qayyum, Zia & Altaf Wasif. 2012. Paraphrase identification using semantic heuristic features. *Research Journal of Applied Sciences, Engineering and Tecnhology* 4(22). 4894–4904.

Zhang, Yitao & Jon Patrick. 2005. Paraphrase identification by text canonicalization. Em *Proceedings of the Australasian Language Technology Workshop*, 160–166.

Zhao, Jiang, Tiantian Zhu & Man Lan. 2014. ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. Em *Proceedings of the International Workshop on Semantic Evaluation*, 271–277.