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

Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual

Overview of the Evaluation of Semantic Similarity and Textual Inference

Erick Rocha Fonseca
Universidade de São Paulo
erickrf@icmc.usp.br

Leandro Borges dos Santos
Universidade de São Paulo
leandrobs@usp.br

Marcelo Criscuolo
Universidade de São Paulo
mcrisc@icmc.usp.br

Sandra Maria Aluísio
Universidade de São Paulo
sandra@icmc.usp.br

Resumo

Inferência Textual e Similaridade Semântica são duas tarefas do processamento de línguas naturais que tratam de pares de trechos de textos. O objetivo da primeira é determinar se o significado de um trecho implica o outro, enquanto que a segunda atribui uma pontuação de similaridade semântica ao par. Esse artigo apresenta os resultados da avaliação conjunta ASSIN (Avaliação de Similaridade Semântica e Inferência) e seu corpus, que foi anotado para ambas as tarefas nas variantes brasileira e europeia da língua portuguesa. O corpus difere de similares na literatura em suas três classes para a tarefa de inferência textual (Implicação, Paráfrase e Neutro) e por ter sido composto de sentenças extraídas de textos jornalísticos. Seis equipes participaram da avaliação conjunta, explorando diferentes estratégias.

Palavras chave

Avaliação conjunta, inferência textual, similaridade semântica

Abstract

Recognizing Textual Entailment and Semantic Textual Similarity are two natural language processing tasks dealing with pairs of text passages. The former aims to determine whether the meaning of one passage entails the other, while the latter assigns a semantic similarity score to the pair. This paper presents the results of the ASSIN shared task and its corpus, annotated for both tasks in the Brazilian and European varieties of the language. The corpus differs from similar ones in the literature in its three RTE classes (Entailment, Paraphrase and Neutral), and for having been composed of sentences extracted from newswire texts. Six teams took part in the shared task, exploring different strategies.

Keywords

Shared task, text entailment, semantic similarity

1 Introdução

A Avaliação de Similaridade Semântica e de Inferência Textual (ASSIN) foi proposta em paralelo com o PROPOR 2016, consistindo em duas subtarefas relacionadas. Ambas as subtarefas dizem respeito ao entendimento de um par de sentenças: a similaridade semântica (STS, *Semantic Textual Similarity*) (Agirre et al., 2015) é uma medida numérica de 1 a 5 do quão similar é o conteúdo das duas sentenças; e a inferência textual (RTE, *Recognizing Textual Entailment*) (Dagan et al., 2013) consiste em classificar o par como tendo uma relação de implicação, paráfrase, ou nenhuma das duas.

A definição exata destas tarefas não é universal. Outros conjuntos de dados apresentam escalas diferentes para a similaridade semântica (Agirre et al., 2015) ou a possibilidade de identificar contradição entre duas sentenças (Bentivogli et al., 2009). No caso do ASSIN, decidimos por uma escala de similaridade de 1 a 5 por achar mais fácil discriminar os diferentes níveis, enquanto na tarefa de inferência, nosso processo de criação de corpus não resultou em quase nenhum caso de contradição.

A avaliação ASSIN 2016 trouxe o primeiro corpus anotado para as duas tarefas em português, incluindo as variantes brasileira e europeia. Foram compiladas sentenças de textos reais, do gênero informativo (textos jornalísticos) em contraste com a abordagem utilizada para a construção de corpora similares em inglês, como SICK (Marelli et al., 2014) e SNLI (Bowman et al., 2015) e dos RTE Challenges (Bentivogli et al., 2009).

Aproveitamos os agrupamentos de notícias por assunto fornecidos pelo *Google News*[1] para

[1] https://news.google.com/

criar o corpus ASSIN 2016. Usamos modelos de espaço vetorial (Turney & Pantel, 2010) para selecionar sentenças similares de documentos diferentes, que passaram por um processo de filtragem manual (onde foram excluídos pares considerados ruidosos) e, por fim, foram anotados por juízes humanos. Cada par foi anotado por quatro pessoas com respeito às duas tarefas.

Participaram do ASSIN seis equipes, sendo três brasileiras e três portuguesas. Cada equipe participante pôde enviar até três saídas dos seus sistemas para cada combinação de variante e subtarefa. As seis equipes participaram da tarefa de similaridade semântica, e quatro delas participaram da inferência textual. É interessante notar que foram exploradas diferentes abordagens para tratar os problemas, mas nem todas foram capazes de superar os *baselines*.

Tratamos brevemente de avaliações conjuntas sobre as mesmas tarefas, para inglês, na Seção 2. Na Seção 3, apresentamos a definição detalhada das tarefas para o escopo do ASSIN 2016. Na Seção 4 descrevemos o processo de criação do corpus, assim como métricas usadas para a avaliação da concordância entre anotadores. Fornecemos também diretrizes para reduzir a subjetividade da anotação. A Seção 5 apresenta as seis equipes participantes e um resumo das suas abordagens. A Seção 6 descreve os *baselines* usados na tarefa e os resultados gerais. As conclusões e possíveis trabalhos futuros são apresentados na Seção 7.

2 Trabalhos Relacionados

A primeira competição de RTE foi o *PASCAL Recognising Textual Entailment Challenge* (RTE-1) (Dagan et al., 2005), que apresentou pares de sentenças coletados manualmente, tentando simular o cenário de aplicações de PLN. Por exemplo, em um cenário de Extração de Informação, a segunda sentença mencionava alguma propriedade de uma entidade mencionada na primeira. Nos anos seguintes, outras edições do evento foram realizadas, trazendo novos corpora anotados. Em particular, no RTE-4 (Giampiccolo et al., 2008), a avaliação trouxe a classificação de alguns pares como contradição. No SemEval 2014, foi utilizado o corpus SICK (Marelli et al., 2014), que trazia anotação tanto de RTE como de STS. Esta foi a última avaliação conjunta para RTE em inglês.

Mais recentemente, foi disponibilizado o corpus SNLI (*Stanford Natural Language Inference*) (Bowman et al., 2015), com cerca de 550 mil pares de sentenças anotados para inferência textual, o maior corpus do gênero até o momento. O SNLI não foi utilizado em nenhuma avaliação conjunta, mas diversos artigos têm sido publicados com experimentos sobre o mesmo, focando normalmente em métodos de *deep learning* (Rocktäschel et al., 2015; Wang & Jiang, 2015). O SNLI e o SICK foram criados a partir de descrições de imagens. No SICK, um processo semi-automático gerou uma segunda sentença para cada descrição, introduzindo negações, trocando palavras, entre outras alterações. Já no SNLI, anotadores escreveram, para cada sentença original, três outras: uma que fosse implicada pela primeira, outra que a contradissesse e uma terceira neutra.

A detecção de similaridade semântica textual foi introduzida em 2012 e, em 2013, foi parte do evento *SEM, acontecendo em conjunto com o SemEval (Agirre et al., 2012, 2013). Desde então, a STS tem sido anualmente uma das tarefas propostas no SemEval. Os pares usados nas avaliações de STS incluem sentenças de diferentes origens, como descrições de vídeos e imagens, manchetes de jornais e diferentes traduções de um mesmo texto.

3 Definição das Tarefas

Apresentamos nessa seção os dois fenômenos anotados no corpus.

3.1 Similaridade semântica

Nossos valores para similaridade semântica variam de 1 a 5, como no corpus SICK, de modo que quanto maior o valor, maior a semelhança do significado das duas sentenças. Esse tipo de medida é inerentemente subjetiva, e não conseguimos chegar a uma definição exata para o que cada valor deveria indicar. Ainda assim, as diretrizes gerais para a pontuação utilizadas no ASSIN 2016 seguem abaixo:

1. As sentenças são completamente diferentes. É possível que elas falem do mesmo fato, mas isso não é visível examinando-as isoladamente, sem contexto.
2. As sentenças se referem a fatos diferentes e não são semelhantes entre si, mas são sobre o mesmo assunto (jogo de futebol, votações, variações cambiais, acidentes, lançamento de produtos).
3. As sentenças têm alguma semelhança entre si, e podem se referir ao mesmo fato ou não.
4. O conteúdo das sentenças é muito semelhante, mas uma (ou ambas) tem alguma informação exclusiva. A diferença pode ser

mencionar uma data, local, quantidade diferente, ou mesmo um sujeito ou objeto diferente.

5. As sentenças têm praticamente o mesmo significado, possivelmente com uma diferença mínima (como um adjetivo que não altera a sua interpretação).

A Tabela 1 mostra exemplos de pares em cada um dos níveis. As diretrizes de anotação requisitavam que se considerasse o conteúdo das sentenças em análise, e não os contextos possíveis nos quais elas poderiam aparecer. Por exemplo, considere o exemplo de similaridade 1 na Tabela 1. Embora seja possível que ambas as sentenças venham do mesmo texto e sejam fortemente relacionadas (o que é o caso nesse exemplo), a anotação não deve considerar essas suposições.

3.2 Inferência Textual

Dagan et al. (2013) definem inferência textual como uma relação unidirecional entre um texto (ou premissa) T e uma hipótese H. Se uma pessoa ao ler T conclui que H é verdadeiro, diz-se que T implica (*entails*) H. Embora seja uma definição subjetiva, ela é largamente aceita na comunidade de processamento de línguas naturais, dada a dificuldade de se chegar a uma definição mais precisa.

É comum a distinção entre pares de textos sem inferência e com contradições em conjuntos de dados de inferência textual. Embora seja interessante a distinção, no corpus ASSIN 2016 eles são raros e dessa forma decidimos não criar uma classe separada. Vale lembrar que, tanto no SICK quanto no SNLI (Bowman et al., 2015), pares com contradição são deliberadamente criados, seja manual ou semi-automaticamente.

Nós também definimos uma classe separada para paráfrases, que embora não sejam frequentes, aparecem em nosso corpus de textos jornalísticos. A Tabela 2 mostra um caso em que a primeira sentença implica a segunda; um caso de implicação mútua ou paráfrase; e um terceiro caso em que não há implicação.

4 Criação do Corpus

Nesta seção descrevemos a criação do corpus e apresentamos as estatísticas da anotação.

4.1 Coleta e Anotação do Corpus

A exploração de agrupamentos de notícias para aquisição de pares de sentenças similares não é uma ideia nova; já foi explorada com sucesso em vários trabalhos da literatura (Dolan et al., 2004; Dagan et al., 2005). Entretanto, em vez de anotadores humanos selecionarem pares com base na sobreposição de palavras, empregamos o *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) para selecionar pares similares.

O LDA, um método de modelagem de espaços vetoriais, atribui uma pontuação para pares de documentos, refletindo o quão similares são entre si. Em um experimento piloto reportado em (Fonseca & Aluísio, 2015), notamos que, em comparação com outros métodos de espaço vetorial, o LDA fornecia os pares mais interessantes para inferência textual, pois recuperava o menor número de sentenças sem relação de inferência (que costumam ser a maioria) e era eficiente em detectar similaridades além da sobreposição de palavras.

Usamos um modelo diferente de LDA para cada variante do português, ambos treinados em grandes corpora de notícias. O modelo para o português do Brasil foi treinado em um corpus coletado do site de notícias G1[2] e o para português europeu com textos do jornal Público[3]. Esses corpora foram somente usados para gerar os modelos LDA, não para coletar os pares de sentenças do corpus ASSIN.

Grupos de notícias sobre o mesmo evento foram coletados do Google News em suas versões específicas para Brasil e Portugal. Filtramos alguns domínios para evitar sites de notícias brasileiros na seção de Portugal e vice-versa. Dados os grupos de notícia coletados e um modelo de espaço vetorial treinado, a criação do nosso corpus seguiu um processo de três etapas:

1. Usamos LDA para encontrar pares de sentenças similares dentro de cada grupo. Esse passo pode ser parametrizado fixando os valores mínimo e máximo de similaridade s_{max} e s_{min}: fixando um valor máximo evita pares de sentenças quase iguais, que seriam classificados trivialmente como paráfrases, e fixando um mínimo evita pares muito dissimilares que são facilmente classificados como sem relação. Fixamos a proporção α de tokens que são encontrados em uma sentença mas não em outra (sem contar *stopwords*). Finalmente, sentenças podem ser limitadas por um tamanho máximo; em uma análise

[2] http://g1.globo.com/
[3] http://www.publico.pt/

1	Mas esta é a primeira vez que um chefe da Igreja Católica usa a palavra em público. A Alemanha reconheceu ontem pela primeira vez o genocídio armênio.
2	Como era esperado, o primeiro tempo foi marcado pelo equilíbrio. No segundo tempo, o panorama da partida não mudou.
3	Houve pelo menos sete mortos, entre os quais um cidadão moçambicano, e 300 pessoas foram detidas. Mais de 300 pessoas foram detidas por participar de atos de vandalismo.
4	A organização criminosa é formada por diversos empresários e por um deputado estadual. Segundo a investigação, diversos empresários e um deputado estadual integram o grupo.
5	Outros 8.869 fizeram a quadra e ganharão R$ 356,43 cada um. Na quadra 8.869 apostadores acertaram, o prêmio é de R$ 356,43 para cada.

Tabela 1: Exemplos para os valores de similaridade semântica.

Inferência	Como não houve acordo, a reunião será retomada nesta terça, a partir das 10h. As partes voltam a se reunir nesta terça, às 10h.
Paráfrase	Vou convocar um congresso extraordinário para me substituir enquanto presidente. Vou organizar um congresso extraordinário para se realizar a minha substituição como presidente.
Sem relação	As apostas podem ser feitas até as 19h (de Brasília). As apostas podem ser feitas em qualquer lotérica do país.

Tabela 2: Exemplos para as categorias de inferência textual.

preliminar, notamos que sentenças muito longas têm muita informação e dificilmente podem ser completamente implicadas por outra.

2. Revisamos os pares coletados em um processo manual. Se um par contém uma sentença sem sentido, é descartado. Sentenças foram também editadas para correção de erros ortográficos e gramaticais, ou para alterar casos em que a presença de implicação é pouco clara.

3. Os pares são anotados. Quatro pessoas anotaram cada par, selecionadas aleatoriamente pelo sistema de anotação. Cada anotador seleciona um valor de similaridade de 1 a 5, e também uma das quatro opções para inferência: a primeira sentença implica a segunda; a segunda implica a primeira; paráfrase, ou nenhuma relação.

Realizamos esse processo em vários lotes, variando os parâmetros. Usamos os valores de s_{min} de 0.65 e 0.6, sem obter grande diferença no resultado. s_{max} foi fixado em 0.9. A proporção de tokens exclusivos para cada sentença foi fixada em 0.1 como mínimo e valores máximos variando entre 0.7 ou 0.8. Com o último valor, notamos um aumento considerável de pares de sentenças com valor de similaridade baixo.

Dada a subjetividade da anotação, definimos algumas diretrizes para lidar com alguns fenômenos linguísticos recorrentes que tinham diferentes interpretações por parte dos anotadores. As diretrizes são voltadas especialmente para a anotação de inferência, e estão listadas na Tabela 3.

Descartamos pares sem concordância de, pelo menos, três votos para a tarefa de inferência textual. Nosso entendimento foi que esses pares eram controversos e assim não seriam boas escolhas para serem incluídos no corpus final. Note-se que os anotadores poderiam indicar implicação tanto da primeira para a segunda sentença como da segunda para a primeira; porém, no corpus final, invertemos a ordem dos pares necessários para que todos os casos de inferência fossem da primeira sentença para a segunda. O valor final de similaridade para cada par é média das quatro pontuações. Dessa forma, os valores são reais separados por intervalos de 0,25.

A anotação foi realizada via uma interface Web construída especialmente para a tarefa, mas flexível o bastante para permitir customizações em futuras anotações. Os anotadores receberam treinamento para calibrar os conceitos das tarefas a serem realizadas, com ajuda de um conjunto de 18 pares exemplificando todos os fenômenos tratados. Em caso de dúvidas, perguntas poderiam ser enviadas via e-mail para a equipe de anotadores, o que permitia discutir casos muito difíceis de decidir, principalmente no começo da anotação.

Por fim, o corpus foi dividido em seções de treinamento (com três mil pares de cada variante) e teste (com os dois mil restantes de cada). A metade brasileira do corpus de treinamento foi disponibilizada em 20 de novembro de 2015, e a metade portuguesa foi disponibilizada dois meses depois.

Conceito	Explicação
Atemporalidade	A interpretação das sentenças não deveria levar em conta a data corrente, de modo que a anotação fizesse sentido no futuro. Assim, embora *há 70 anos atrás* e *em 1945* sejam equivalentes em 2015, devem ser considerados distintos pelos anotadores.
Entidades Nomeadas	Entidades nomeadas que aparecem nas duas sentenças, tendo um aposto ou adjetivo em uma delas, devem ser consideradas equivalentes. *Florianópolis, em Santa Catarina* é equivalente a apenas *Florianópolis*.
Discurso Indireto	Uma sentença com discurso indireto (i.e., *O embaixador disse que (...)*) pode implicar outra que contenha apenas a fala atribuída. O contrário, no entanto, não é possível.
Quantidades	Valores numéricos diferentes só podem ser aceitos para paráfrase/implicação se tiverem indicadores explícitos de serem aproximações: *acerca de*, *pelo menos*, *quase*, *perto de*, etc. Por exemplo, *arrecadou 7 milhões* não implica arrecadou *6 milhões* pois, mesmo sendo uma quantia menor, é possível que se refira a outro evento.

Tabela 3: Resumo das Diretrizes para Anotação.

4.2 Estatísticas da Anotação

O corpus foi anotado por 36 pessoas, que participaram em diferentes quantidades: o anotador com menor participação julgou 25 pares, enquanto o com maior participação julgou 6.740.

Do total de pares anotados, 11.3% foram descartados por não terem três julgamentos iguais quanto à implicação. A proporção é um pouco menor do que as reportadas na criação dos corpora RTE Challenge (Dagan et al., 2005; Giampiccolo et al., 2007). No total, o ASSIN tem 10 mil pares, sendo metade em português brasileiro e metade em português europeu.

A Tabela 4 sumariza estatísticas da anotação. A correlação ρ de Pearson é uma boa métrica para a concordância entre anotadores (ou para o desempenho de um sistema), tendo sido usada também pelos organizadores das competições de STS. Essa medida avalia a dependência linear entre duas variáveis, o que é mais informativo do que apenas a correlação de ranqueamento (computável com a correlação de Spearman). Por exemplo, se um anotador avalia três pares com semelhança 2, 3 e 4, enquanto outro avalia os mesmos com 2, 4 e 5, o ranqueamento é idêntico, mas o valor de ρ está abaixo de 1 por não serem duas variáveis (perfeitamente) linearmente dependentes.

O valor de ρ apresentado na tabela se refere à média das correlações calculadas entre todos os anotadores, ponderada pela quantidade de pares que cada um anotou. Para cada anotador, calculamos a correlação das suas pontuações de similaridade com as médias das pontuações dos pares que ele ou ela anotou (excluindo a sua anotação do cômputo). Para efeito de comparação, a anotação do STS 2015 obteve valores entre 0.65 e 0.85, o que mostra que alcançamos boa concordância entre anotadores quanto à similaridade.

Métrica	Valor
Correlação de Pearson	0,74
Desvio Padrão Médio	0,49
κ de Fleiss	0,61
Concordância	0,80

Tabela 4: Estatísticas da Anotação. Os primeiros 2 valores se referem à anotação de similaridade; os 2 últimos valores à inferência.

O desvio padrão médio avalia a divergência dos julgamentos de similaridade dos pares. É calculado como a média dos desvios padrão de todos os pares no corpus; esses, por sua vez, são calculados como o desvio padrão das quatro pontuações em relação à média do par. O valor reportado na anotação do SICK é de 0,76, indicando que as pontuações dos nossos anotadores divergiram menos.

Com relação à inferência, o valor da concordância κ de Fleiss foi relativamente baixo, o que indica que a anotação desta tarefa de fato envolveu boa quantidade de subjetividade. Os corpora dos desafios RTE, por exemplo, tiveram uma taxa de concordância maior: 0,6 na primeira edição (Dagan et al., 2005), mas chegando a 0,75 ou mais nas subsequentes (Giampiccolo et al., 2007). Entretanto, deve ser notado que esses corpora tratam de sentenças curtas como segundo componente do par (a sentença implicada), o que torna a decisão mais fácil.

A última linha da tabela se refere à concordância simples entre os anotadores. Isso significa que, em 80% dos casos, dois anotadores que julgaram o mesmo par escolheram a mesma categoria de inferência.

As tabelas 5 e 6 mostram estatísticas sobre as anotações de similaridade e inferência, respectivamente. Pode-se ver que as pontuações de si-

milaridade mais comuns estão no intervalo entre 2 e 3. Já quanto à inferência, percebe-se que a relação neutra é a classe majoritária, enquanto as paráfrases são uma porção pequena do corpus.

Similaridade	**PB**	**PE**	**Total**
4,0 – 5,00	1.074	1.336	2.410
3,0 – 3,75	1.591	1.281	2.872
2,0 – 2,75	1.986	1.828	3.814
1,0 – 1,75	349	555	904
Média	3,05	3,05	3,05

Tabela 5: Estatísticas de similaridade do ASSIN.

Relação	**PB**	**PE**	**Total**
Sem relação	3.884	3.432	7.316
Implicação	870	1.210	2.080
Paráfrase	246	358	604

Tabela 6: Estatísticas de inferência do ASSIN.

A pouca quantidade de pares com relação de inferência foi notada já durante nossa análise de um corpus piloto, que não foi incluído no corpus final. Isso se devia ao fato de que, em muitos casos, apenas alguns detalhes impediam que houvesse tal relação: a menção a um local, tempo, propósito, entre outros. Essa situação é ilustrada no exemplo a seguir.

(1) a. O Internacional manteve a boa fase e venceu o Strongest por 1 a 0 nesta quarta-feira, garantindo a liderança do Grupo 4 da Libertadores.

b. Em casa, a equipe gaúcha derrotou o The Strongest, por 1 a 0, e garantiu a primeira colocação do Grupo 4 da Copa Libertadores.

Apesar de as duas sentenças compartilharem a maior parte do conteúdo, cada uma tem alguma informação específica que não é implicada pela outra. A primeira menciona o nome da equipe, além de que estava em boa fase e que o jogo foi na quarta-feira. Já a segunda diz que o jogo foi na casa da equipe, sem explicitar seu nome. Esse tipo de fenômeno é particularmente comum quando se tratam de sentenças longas.

Visando aumentar a proporção de pares com inferência, realizamos pequenas mudanças nas sentenças durante a segunda etapa do nosso processo listado na Seção 4.1. Assim, passamos a remover pequenos trechos de uma ou ambas as sentenças, caso as alterações possibilitassem a inferência. Apesar da proporção final estar menos desequilibrada que o observado em nosso corpus piloto, ainda temos menos pares com inferência e especialmente paráfrases do que o que gostaríamos.

5 Sistemas Participantes

Participaram do ASSIN seis equipes, sendo três brasileiras e três portuguesas. Cada equipe participante pôde enviar o resultado de até três execuções de seus sistemas para cada combinação de variante da língua e subtarefa.

Na tarefa de similaridade, participaram todas as seis equipes inscritas, enquanto quatro participaram da tarefa de inferência textual. A L2F/INESC-ID foi a única a reportar resultados apenas para uma variante; no caso, o português europeu[4].

É interessante notar que os participantes adotaram estratégias bastante diversas entre si, o que permite uma análise de diferentes pontos de vista sobre as tarefas. Ressaltamos também que as equipes que participaram de ambas as tarefas usaram os mesmos atributos para treinar diferentes modelos (em alguns casos, com uma etapa intermediária de seleção automática de atributos).

Portanto, não fazemos aqui uma separação entre abordagens específicas de cada subtarefa; em vez disso, resumimos brevemente o funcionamento dos sistemas dos participantes a seguir.

5.1 Abordagens

A equipe Solo Queue (Hartmann, 2016) utilizou uma abordagem bastante simples, baseada apenas no valor da similaridade do cosseno de duas representações vetoriais de cada sentença. Tais representações são geradas como a soma dos vetores de cada palavra, que por sua vez são obtidas por meio de TF-IDF e word2vec (Mikolov et al., 2013).

Em seguida, os cossenos entre as duas representações (TF-IDF e word2vec) de cada sentença são dadas como entrada para um regressor linear que determina a similaridade do par.

O sistema de L2F/INESC-ID (Fialho et al., 2016) consistiu em extrair diversas métricas dos pares de sentenças, como distância de edição, palavras em comum (incluindo métricas separadas para entidades nomeadas ou verbos modais),

[4]Os autores informaram que não houve tempo o suficiente para treinar os seus modelos para o português do Brasil antes do prazo da avaliação conjunta. Ainda assim, apresentam em seu artigo resultados obtidos após a data.

BLEU, ROUGE etc. Tais métricas foram computadas tanto das sentenças originais como de outras versões, que poderiam estar em caixa baixa, com palavras radicalizadas, usando clusters de palavras (Turian et al., 2010), entre outras modificações. A combinação de diferentes versões das sentenças com as diferentes métricas gerou mais de 90 atributos para descrever cada par, que são então usados para treinar um Kernel Ridge Regression (para similaridade) e um SVM (para inferência).

Fialho et al. (2016) experimentaram ainda aumentar o conjunto de treinamento com uma versão do corpus SICK traduzida automaticamente para o português. No entanto, os resultados obtidos ao se treinar o regressor na versão aumentada foram inferiores, provavelmente devido aos erros de tradução. Por fim, os autores avaliam seus modelos quando treinados em uma variante do português e testados na outra.

As equipes ASAPP e Reciclagem (Alves et al., 2016) compartilharam um módulo de análises de relações lexicais baseado em redes semânticas (como tesauros e wordnets). Diversas métricas baseadas em tais relações foram extraídas dessas redes.

O Reciclagem não conta com nenhum módulo de aprendizado de máquina, empregando apenas métricas de similaridade baseadas nas relações semânticas entre as palavras das duas sentenças. Nesse sentido, o método teve um caráter exploratório quanto à capacidade de diferentes redes semânticas contribuírem para a tarefa de STS e do quanto um sistema sem treinamento poderia alcançar em termos de performance.

Já o ASAPP emprega, além das métricas usadas pelo Reciclagem, atributos como contagem de tokens de cada sentença, orações nominais, tipos de entidades nomeadas etc., todos dados como entrada para classificadores e regressores. Em suas três execuções, foram exploradas formas de partição de dados, combinação de modelos e redução da quantidade de atributos.

Barbosa et al. (2016) utilizaram a estratégia proposta por Kenter & de Rijke (2015): são obtidas representações vetoriais das palavras (no caso, foi usado o word2vec) e, em seguida, os vetores de uma sentença são comparados com os da outra, obtendo-se medidas baseadas no cosseno e a distância euclidiana.

Todas as medidas obtidas são então agrupadas em histogramas, com intervalos pré-definidos. São usados diferentes histogramas para cada medida, e as suas contagens são dados como entrada para os modelos de aprendizado de máquina. Para a tarefa de similaridade, foram usados SVR e o método Lasso, e para a inferência, apenas um SVM.

Também foram explorados métodos baseados em redes neurais recorrentes e convolucionais, usando uma arquitetura siamesa. Esse tipo de arquitetura usa o mesmo conjunto de pesos para mapear cada uma das sentenças para um vetor. Dados os dois vetores, pode ser calculado diretamente o seu cosseno, que é então mapeado para um valor de similaridade. No entanto, a despeito dos bons resultados reportados na literatura recente em PLN, as redes neurais obtiveram resultados muito abaixo dos outros métodos usados pela equipe. A provável causa desta desparidade é a quantidade relativamente pequena de dados disponíveis no ASSIN.

A equipe FlexSTS (Freire et al., 2016) apresentou um framework para calcular a similaridade semântica textual baseada em combinar medidas de semelhança entre tokens de acordo com alinhamentos entre eles. Foram exploradas três configurações: a primeira treina um regressor usando apenas uma função DICE e medidas de distâncias entre os tokens na WordNet. Foi usada a WordNet da língua inglesa, e os pares do ASSIN foram traduzidos automaticamente para consultá-la.

A segunda abordagem do FlexSTS usou apenas o modelo HAL (Hyperspace Analogue to Language) para calcular a similaridade entre as palavras mais similares, enquanto a terceira abordagem combina o modelo HAL com a WordNet. Essas duas não usam nenhum componente de aprendizado de máquina, recorrendo a fórmulas pré-definidas para computar o valor de similaridade de cada par.

6 Avaliação e Resultados

Os participantes receberam o conjunto de teste (sem os rótulos corretos dos pares) em 4 de março de 2016, e tiveram 8 dias para enviar aos organizadores os arquivos com as respostas produzidas por seus sistemas. Cada participante pôde enviar até três resultados.

As métricas usadas na avaliação das duas tarefas são consoantes com as usadas em avaliações conjuntas internacionais. Na tarefa de similaridade textual, foi usada a correlação de Pearson, tendo o erro quadrático médio (MSE, *mean square error*) como medida secundária. Idealmente, os sistemas devem ter a maior correlação possível e o menor MSE possível. Para a inferência, foi usada a medida F1, tendo a acurácia como medida secundária.

6.1 Baselines

Foram usadas duas estratégias como *baseline* para o ASSIN: a primeira memoriza a média das similaridades do corpus de treino e a classe de inferência mais comum, e emite esses valores para todos os pares de teste. A segunda, um pouco mais sofisticada, consiste no treinamento de um classificador baseado em regressão logística e um regressor linear. Estes dois modelos são treinados com apenas dois atributos: a proporção de tokens exclusivos da primeira e da segunda sentença.

6.2 Resultados

As Tabelas 7 e 8 listam os resultados das tarefas de similaridade e inferência, respectivamente, obtidos por cada participante em suas três execuções, bem como os resultados dos sistemas *baseline*.

A equipe Solo Queue (Hartmann, 2016) obteve os melhores resultados da similaridade semântica para o português do Brasil, enquanto o Blue Man Group (Barbosa et al., 2016) obteve os melhores resultados para inferência textual . Já com o português europeu, a L2F/INESC-ID (Fialho et al., 2016) alcançou os melhores resultados nas duas tarefas.

O primeiro *baseline* obteve 0 na correlação de Pearson pelo fato de não haver variação em suas respostas, e a medida ser baseada na correlação de duas variáveis. Ao se combinar as respostas para as duas metades do corpus, é obtido um valor negativo, indicando uma performance pior que dar a mesma resposta sempre.

No entanto, considerando o MSE, esse *baseline* teve resultados melhores que algumas execuções dos participantes, o que significa que tais execuções computaram valores muito distantes da similaridade real dos pares. Já o segundo *baseline* teve resultados competitivos, chegando a superar diversas execuções.

Quanto à inferência, com resultados na Tabela 8, o primeiro *baseline* é também facilmente superado, mas o segundo se saiu bastante bem. Na variante brasileira, chegou a superar todos os três participantes e, na europeia, apenas uma execução da L2F/INESC-ID se saiu melhor.

O último resultado foi bastante inesperado. Apesar de toda a modelagem do problema feita pelas equipes participantes, um *baseline* com apenas dois atributos simples, sem acesso a nenhum recurso externo e usando apenas modelos lineares foi capaz de superar quase todos os sistemas empregados na tarefa. Ao mesmo tempo, esse resultado indica que a presença de inferência no ASSIN é fortemente relacionada com a sobreposição lexical, ainda que tenhamos nos esforçado em incluir tanto pares com inferência que tivessem palavras distintas quanto pares sem relação e palavras compartilhadas.

7 Conclusões

Descrevemos a proposta da Avaliação de Similaridade Semântica e Inferência Textual, como foi criado seu corpus anotado, quais foram as equipes participantes da avaliação conjunta e os resultados que obtiveram. Apresentamos, ainda, dois sistemas *baseline* bastante simples, mas dos quais um superou a maioria dos participantes na tarefa de inferência textual.

O ASSIN 2016 cumpriu seu objetivo de trazer a primeira avaliação conjunta de inferência textual e similaridade semântica para o português. Listamos a seguir algumas conclusões que dizem respeito à criação do corpus e aos sistemas desenvolvidos para a tarefa.

7.1 Criação do Corpus

O método que usamos para a compilação do corpus, apesar de funcional, apresenta alguns empecilhos. O primeiro é o gargalo da etapa de limpeza antes da anotação em si. Durante essa etapa, os critérios para se eliminar ou editar pares são bastante delicados, como nossa experiência mostrou. É uma parte da anotação que deve ficar a cargo de pessoas que tenham conhecimento sobre a tarefa e seus objetivos, e dificilmente poderia ser delegada para uma plataforma de *crowdsourcing*.

Outra dificuldade diz respeito à subjetividade da tarefa. Em alguns casos, os anotadores gastaram bastante tempo tentando se decidir quanto aos julgamentos que deveriam dar para certos pares. Esse tipo de problema retoma o anterior: certas alterações no conteúdo das sentenças torna as decisões mais fáceis, e portanto, a anotação mais confiável e produtiva.

7.2 Sistemas Participantes

Os participantes do ASSIN exploraram diferentes tipos de estratégia para as duas tarefas propostas. É particularmente interessante notar que dentre os melhores resultados obtidos estão duas abordagens muito simples: na similaridade semântica, a comparação da combinação de vetores de palavras, como feito pelo Solo Queue; e para inferência, a comparação da proporção de

		PB		PE		Geral	
Equipe	**Exec.**	**Pearson**	**MSE**	**Pearson**	**MSE**	**Pearson**	**MSE**
Solo Queue	1	0,58	0,50	0,55	0,83	0,56	0,66
	2	0,68	0,41	0,00	1,55	0,29	0,98
	3	**0,70**	**0,38**	0,70	0,66	**0,68**	**0,52**
Reciclagem	1	0,59	1,36	0,54	1,10	0,53	1,23
	2	0,59	1,31	0,53	1,14	0,54	1,23
	3	0,58	1,37	0,53	1,18	0,53	1,27
Blue Man Group	1	0,65	0,44	0,63	0,73	0,63	0,59
	2	0,64	0,45	0,64	0,72	0,63	0,59
ASAPP	1	0,65	0,44	0,68	0,70	0,65	0,57
	2	0,65	0,44	0,67	0,71	0,64	0,58
	3	0,65	0,44	0,68	0,73	0,65	0,58
LEC-UNIFOR	1	0,62	0,47	0,64	0,72	0,62	0,59
	2	0,56	2,83	0,59	2,49	0,57	2,66
	3	0,61	1,29	0,63	1,04	0,61	1,17
L2F/INESC-ID	1			**0,73**	**0,61**		
	2			0,63	0,70		
	3			0,63	0,70		
Baseline (média)	–	0,00	0,76	0,00	1,19	-0,08	0,97
Baseline (sobreposição)	–	0,63	0,46	0,64	0,75	0,62	0,60

Tabela 7: Resultados de todas as execuções para a tarefa de similaridade semântica.

		PB		PE		Geral	
Equipe	**Exec.**	**Acurácia**	**F1**	**Acurácia**	**F1**	**Acurácia**	**F1**
Reciclagem	1	77,65%	0,29	73,10%	0,43	75,38%	0,40
	2	79,05%	0,39	72,10%	0,38	75,58%	0,38
	3	78,30%	0,33	70,80%	0,32	74,55%	0,32
Blue Man Group	2	81,65%	0,52	77,60%	0,61	79,62%	0,58
ASAPP	1	81,20%	0,50	77,75%	0,57	79,47%	0,54
	2	81,65%	0,47	78,90%	0,58	80,27%	0,54
	3	77,10%	0,50	74,35%	0,59	75,72%	0,55
L2F/INESC-ID	1			**83,85%**	**0,70**		
	2			78,50%	0,58		
	3			78,50%	0,58		
Baseline (maioria)	–	77,65%	0,29	69.30%	0,27	73,47%	0,28
Baseline (sobreposição)	–	**82,80%**	**0,64**	81,75%	**0,70**	**82,27%**	**0,67**

Tabela 8: Resultados de todas as execuções para a tarefa de inferência textual.

palavras exclusivas de cada sentença, que foi um dos *baselines* propostos.

Todavia, a equipe L2F/INESC-ID obteve os melhores resultados do ASSIN na variante europeia (a única em que competiu), empregando um sistema baseado em um rico conjunto de atributos. Esse resultado indica que superar métodos simples como os listados acima requer uma modelagem extensiva do problema.

Outra linha de pesquisa bastante bem sucedida na literatura recente são redes neurais recorrentes (como LSTMs) ou convolucionais. O Blue Man Group foi o único grupo a explorá-las, mas as descartou após obter resultados preliminares negativos. Uma possível explicação para esse fato é que o conjunto de dados do ASSIN é menor e com sentenças mais complexas do que as que se encontram para conjuntos semelhantes

em inglês, onde os modelos neurais obtêm os melhores resultados.

Por fim, notamos que nenhum dos participantes modelou as sentenças em alguma estrutura sintática ou semântica; em vez disso, todos exploraram apenas o nível lexical. Pelo menos para a inferência textual, há evidências na literatura de que a compreensão da estrutura das sentenças tem um papel importante (Dagan et al., 2013), e a ausência desse tipo de análise pode explicar o desempenho dos sistemas abaixo do *baseline*.

7.3 Trabalhos Futuros

Novas edições do ASSIN teriam o potencial de estimular e melhorar a pesquisa nas duas tarefas propostas para a língua portuguesa. No entanto, acreditamos que seria interessante trabalhar com outros tipos de pares de sentença, especialmente na tarefa de inferência.

Uma possibilidade seria o uso de pares de sentenças escritos especificamente com o objetivo de terem ou não uma relação de implicação, como foi feito no SICK e SNLI. Nesse caso, a subjetividade da anotação é reduzida drasticamente, com o preço de não se trabalhar com um cenário realista. De fato, a motivação principal da criação destes dois corpora foi fornecer um ambiente para sistemas de PLN aprenderem o funcionamento de certos mecanismos da linguagem humana.

Outro direcionamento seria usar apenas fatos simples, na forma de sentenças com uma única oração, como o segundo componente de cada par. Essa foi a estratégia adotada na criação dos corpora dos RTE Challenges, e mantêm o realismo da tarefa na medida em que a primeira sentença pode ser extraída de um jornal ou outra fonte real. Por outro lado, esse cenário não requer que os sistemas processem e comparem duas sentenças inteiras, mas apenas busque por confirmação de um fato.

Por fim, uma estratégia que facilitasse a anotação do corpus seria também interessante por permitir a criação um novo recurso em maior escala, tornando mais viável a exploração de métodos neurais que necessitam de grandes bases de treinamento.

Agradecimentos

Agradecemos o apoio da Fapesp, processos número 2016/02466-5 e 2013/22973-0, o apoio do CNPq, processos número 155137/2015-8 e 153047/2016-0, e também o apoio da Google via programa *Google Research Awards for Latin America*, projeto 23327 Google/FUNDEP Google Research Grant para o desenvolvimento dessa pesquisa.

Referências

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona 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. Em *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 252–263.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre & Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. Em **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics.*, 32–43. Association for Computational Linguistics.

Agirre, Eneko, Daniel M. Cer, Mona T. Diab & Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. Em *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, 385–393.

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

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.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo & Bernardo Magnini. 2009. The fifth Pascal recognizing textual entailment challenge. Em *Proceedings of the Text Analysis Conference 2009*, s.pp.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bowman, Samuel R., Gabor Angeli, Christopher Potts & Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. Em *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. ACL.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. Em *Proceedings of the PASCAL challenges on Recognizing Textual Entailment*, 177–190.

Dagan, Ido, Dan Roth, Mark Sammons & Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications* Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Dolan, Bill, Chris Quirk & Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. Em *Proceedings of the 20th International Conference on Computational Linguistics*, 350–356.

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 R. & Sandra M. Aluísio. 2015. Semi-Automatic Construction of a Textual Entailment Dataset: Selecting Candidates with Vector Space Models. Em *Proceedings of STIL 2015*, 201–210.

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

Giampiccolo, Danilo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio & Bill Dolan. 2008. The fourth PASCAL recognizing textual entailment challenge. Em *Proceedings of the First Text Analysis Conference*, 1–9.

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan & Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. Em *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, 1–9.

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

Kenter, Tom & Maarten de Rijke. 2015. Short text similarity with word embeddings. Em *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1411–1420.

Marelli, Marco, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini & Roberto Zamparelli. 2014. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Em *Proceedings of the 8th International Workshop on Semantic Evaluation*, 1–8.

Mikolov, Tomas, Kai Chen, eg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Available from arXiv:1301.3781.

Rocktäschel, Tim, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský & Phil Blunsom. 2015. Reasoning about entailment with neural attention. Available from arXiv:1509.06664.

Turian, Joseph, Lev Ratinov & Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. Em *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.

Wang, Shuohang & Jing Jiang. 2015. Learning natural language inference with LSTM. Available from arXiv:1512.08849.