http://dxdoi.org/10.11606/issn.2317-9511.v30i0p71-86

Validação de termos de domínio por meio de uma base lexical-semântica difusa

Domain terms validation by means of a fuzzy lexical-semantic base

Afonso Xavier Canosa Rodrigues[1]

Resumo: A extração ou reconhecimento de termos pesquisa um *corpus* para prover uma lista de termos específicos de domínio a fim de ser usada em trabalhos mais avançados tais como a construção de terminologias e ontologias. Tanto medidas estatísticas quanto técnicas do Processamento da Linguagem Natural (PLN) têm sido investigadas para melhorar o desempenho na precisão das listas recuperadas. Não obstante, para manter a abrangência alta, as listas contêm falsos positivos. Para validar os candidatos como verdadeiros positivos, os termos têm de ser avaliados quer manualmente, quer automaticamente, por contraste com recursos externos, nomeadamente glossários específicos. Apresentamos uma série de experiências que mostram como uma base de conhecimento lexical pode melhorar o desempenho destes glossários de modo significativo. Partimos de uma lista de 50 candidatos a termos de domínio com precisão de 52%. Por meio da uma base lexical difusa, em que as palavras são agrupadas com um valor de associação semântica, achamos valores de corte para atingir percentagens de 100% tanto na precisão quanto na abrangência sobre a lista de partida, mantendo o valor da medida-F > 80%, com melhor resultado em 90%. Concluímos que, considerando que é necessário mais trabalho na pesquisa de limites e diferentes cenários, uma base lexical difusa pode melhorar o estado da arte das abordagens convencionais da extração automática de termos.

Palavras-chave: extração automática de termos; relações semânticas; synsets difusos.

[1]Licenciado em Filologia Galego-Portuguesa e investigador do programa de doutoramento de Geografia da Universidade de Santiago de Compostela.

Abstract: Term extraction or recognition searches a given corpus to provide a list of domain specific terms for further use in more advanced tasks as in terminology and ontology building. Several statistical measures and Natural Language Processing techniques have been researched to improve precision of retrieved lists. However, to keep recall high, lists contain a number of false positives. To validate candidates as true positives in the domain, terms have to be manually evaluated or automatically checked against external resources such as specialized glossaries. Starting with a baseline of 50 candidate terms with 52% precision, we perform a series of experiments to show that a lexical knowledge base can significantly improve glossary performance. Furthermore, using a fuzzy lexical base, words clustered by a semantic association value, we research cutting points to reach 100% rates for either precision or recall for the baseline list, while keeping F-Measure > 80%, achieving 90% as best result. We conclude that, considering further research for limits and different case scenarios is also needed, a fuzzy lexical base can improve current state-of-the art approaches in automatic term extraction .

Keywords: automatic term extraction; semantic relations; fuzzy synsets.

1. Introdução

Apresentamos um método para a melhoria dos resultados de extração de termos de domínio pela aplicação de uma base lexical-semântica difusa, o CLIP 2.1, um recurso lexical que organiza termos em relação de sinonímia outorgando um valor numérico para o grau de pertença de cada termo num conjunto (chamado de synset, seguindo o modelo de uma Wordnet). Através de uma lista de termos mais representativos de um documento, obtida por meio de técnicas convencionais de extração de termos, queremos melhorar os resultados de precisão sem reduzirmos a abrangência. Este processo pode ser também referido como validação, resolvida mais frequentemente quer manualmente, quer pela aplicação de listas especializadas. Neste último caso, temos o problema de que, ainda assumindo que as listas tenham uma precisão de 100%, não abrangem necessariamente todos os termos do domínio. Para superar esta dificuldade propomos o uso de uma base de conhecimento lexical. Analisamos uma série de experiências em que glossários geográficos servem de listas semente para obtermos *synsets* do domínio da base lexical difusa CLIP 2.1, de onde avaliamos, considerando uma métrica de associação

semântica difusa, uma lista de termos candidatos extraídos de um *corpus* do domínio geográfico. Obtemos resultados > 80% na medida-F[2] sem reduzirmos a abrangência atingida pelos métodos convencionais estatísticos e de Processamento da Linguagem Natural (PLN). O melhor resultado ao otimizar tanto a precisão quanto a abrangência atinge 90% na medida-F com uma abrangência > 80%.

2. Extração de termos de domínio e bases lexicais

Termos de um domínio são aqueles que conformam a terminologia específica de uma área. A extração de termos num *corpus* visa obter listas de uma temática e aplica critérios de seleção para reduzir o número de candidatos (CONRADO; FELIPPO; PARDO; REZENDE 2014). As abordagens de reconhecimento automático de termos aplicam métodos estatísticos e atributos linguísticos, a soma de ambos propicia modelos híbridos (CONRADO; PARDO; REZENDE 2015). Como resultado obtemos um vocabulário que finalmente tem de ser validado por especialistas do domínio (ALMEIDA, ALUÍSIO, OLIVEIRA 2007) ou contrastado com listas especializadas e recursos externos (WENDT; LOPES; MARTINS; VIEIRA; LIMA 2010).

A extração de léxicos preferenciais (ZAPPAROLI 2010) ligados à temática de um conjunto de textos *tem* sido atendida na linguística de *corpus* a partir da comparação das frequências normalizadas do léxico contido em vários documentos de que se capturam os vocábulos com maior contraste. Lopes, Fernandes e Vieira (2016) comparam o efeito das métricas de base estatística e desenvolvem uma variante específica da medida TF-IDF (*Term Frequency Inverse Document Frequency*) para a recuperação de termos, partindo também de um princípio contrastante, mas centrando o trabalho na extração

[2]Medida relativamente à precisão e abrangência segundo a fórmula
F =2 x Precisão x Abrangência / (Precisão + Abrangência)

de termos, ainda que limitando a análise a bigramas e trigramas que, de início, têm já uma precisão > 80%. Teixeira (2011) analisou o desempenho de ferramentas automáticas para o ciclo completo de extração de termos, achando os melhores desempenhos ainda muito baixos, na faixa de 20 % para o índice de verdadeiros positivos. Como parte de uma proposta de aprendizagem máquina, Conrado (2014) oferece uma ampla visão de conjunto das métricas e técnicas de extração de terminologia baseada em *corpora* de domínio para o português.

A proposta que analisamos neste artigo é aplicável a qualquer lista de termos de um domínio, independentemente da métrica considerada e do número de *corpora* usados. Dada a lista a avaliar, procedemos a validá-la pela aplicação de recursos lexicais. Uma base de dados lexical organiza vocábulos segundo de um modo similar a um tesouro lexicográfico, mostrando relações semânticas e agrupando termos de um domínio ou área temática. Um modelo de grande aplicação é a WordNet (FELLBAUM 1998), que serve mais frequentemente de modelo para a criação de bases de conhecimento lexical em Português (GONÇALO OLIVEIRA; GOMES 2010; FELIPPO; ALMEIDA 2010, 2014; GONÇALO OLIVEIRA; PAIVA; FREITAS; RADEMAKER; REAL; SIMÕES 2015). A particularidade da base lexical usada nesta proposta é ela ser difusa, isto é, os termos do synset vêm ordenados segundo uma métrica que determina o grau de associação semântica ao termo alvo de entrada.

3. Materiais

Tanto os materiais no seu estado inicial quanto as ferramentas para o seu processamento são preferentemente de acesso aberto.

Corpus do domínio geográfico

Partimos de um *corpus* do domínio geográfico com as entidades geográficas mencionadas já anotadas. O documento inicial é uma edição

digital da *Peregrinação*[3] semi-normalizada para o padrão atual com scripts da nossa elaboração. O projeto formou parte da nossa tese de doutoramento, que tinha como objetivo indexar e georreferenciar todos os topônimos contidos na obra de Fernão Mendes Pinto. A obtenção de termos de domínio geográfico serviu para descrever e classificar os topônimos. Os termos a extrair são, portanto, nomes comuns (não entidades) do domínio geográfico. Partindo da observação de que existem termos geográficos que aparecem frequentemente perto das entidades geográficas mencionadas, selecionamos aquelas orações que contêm como mínimo uma entidade geográfica mencionada para criamos um *subcorpus* de âmbito geoespacial.

Glossários especializados

Selecionamos dois glossários como listas especializadas, o IBGE (IBGE, 2015) com 126 termos geográficos usados no mapeamento contemporâneo do Brasil e o listado de termos de domínio da taxonomia de GeoNames[4], originalmente em inglês e traduzida para o português mediante o sistema de tradução automática de Google[5] com o resultado de 667 termos. Da sua união obtemos um terceiro glossário, IBGE ∪ GeoNames_trad, que contém 725 termos, os comuns e não comuns de IBGE e a tradução de GeoNames.

Base lexical difusa

Como terceiro recurso usamos o CLIP 2.1[6] (GONÇALO OLIVEIRA; GOMES 2014; SANTOS; GONÇALO OLIVEIRA 2015), uma base lexical organizada em *clusters* como os *synsets* em uma WordNet, com a diferença de que os termos aparecem agrupados de modo difuso, com um valor numérico determinado pelo cálculo de coocorrências a partir da análise de bases lexicais de relações de sinonímia (SANTOS; GONÇALO OLIVEIRA 2015).

cidade(1.7777778); urbe(1.0); metrópole(0.7777778); município(0.5); cidade-estado(0.44444445); centro_urbano(0.44444445); cidades-estado(0.44444445); cidade-

[3] Primeira edição fac-similar disponível em http://purl.pt/82
[4] http://www.GeoNames.org/export/codes.html
[5] https://translate.google.com/?hl=pt-PT
[6] http://ontopt.dei.uc.pt/index.php?sec=contopt

livre(0.44444445); praça(0.4); metrópoles(0.33333334); cividade(0.22222222); empório(0.2); concelho(0.2); foco(0.1); chia(0.1); aldeias(0.1); autarquias(0.1); povoado(0.1); comuna(0.1); aldeia(0.1); arraial(0.1); nação(0.1); centro(0.1); cabeça(0.1); morada(0.1); capital(0.1); burgo(0.1); vila(0.1);

Tabela 1: Exemplo de *synset* no CLIP 2.1. Os termos agrupados contêm um valor de associação semântica (entre parêntese).

4. Procedimento

4.1 Obtenção da lista de candidatos a termos de domínio

A lista a validar foi obtida do melhor resultado de uma combinação de métodos de extração de termos com a métrica estatística TF-IDF (SALTON; BUCKLEY 1988) e técnicas de PLN sobre o *subcorpus* elaborado a partir das orações que continham como mínimo uma entidade geográfica mencionada anotada. Os termos candidatos foram sucessivamente reduzidos pela aplicação de janelas de n-gramas com a entidade geográfica mencionada como centro e filtrados sintaticamente pela anotação da categoria gramatical. Os resultados de PLN foram melhorados com fórmulas TF-IDF segundo o sistema SMART (SINGHAL; SALTON; BUCKLEY 1996) com normalizações pelo cosseno sobre a medida da frequência absoluta e com mais uma variante sobre a frequência relativa. Todas as métricas foram implementadas em scripts com o software estatístico de R[7]. Para a obtenção de matrizes de coocorrências, aplicação das listas filtros e comparação de resultados das métricas TF-IDF, usamos o pacote TM (FEINERER; HORNIK; MEYER 2008). Obtivemos, deste modo, listas de 50 termos candidatos a termos de domínio. A Tabela 2 mostra o melhor resultado que chamamos lista a validar.

cidade, ***reino***, ***ilha***, ***porto***, ***rio***, *nome*, ***lugar***, ***fortaleza***, ***terra***, ***costa***, *dia*, ***barra***, ***enseada***, *ano*, *capitão*, *tempo*, ***vila***, *filho*, *gente*, ***casa***, ***estreito***, ***império***, ***mar***, *partes*, *dom*, ***estado***, *morte*, *naos*, *parte*, ***povoação***, *armada*, *mercador*, *nao*, ***fazenda***, *padre*, *senhor*, ***serra***, ***caminho***, ***campo***, *embaixador*, *junco*, ***lago***, *legoas*, *viagem*, *cousa*, *guerra*, ***monte***, *partido*, ***ponta***, ***castelo***.

Tabela 2: Lista de 50 termos candidatos extraídos do *subcorpus* do domínio geográfico (lista a validar). Em negrito os termos avaliados manualmente como verdadeiros positivos.

[7] http://www.R-project.Org

A lista de candidatos foi avaliada primeiro manualmente (termos em negrito na Tabela 2), obtendo assim uma precisão de 52%. O objetivo das experiências é validar automaticamente a lista a fim de aumentar a precisão sem diminuir o número de verdadeiros positivos. Tanto a abrangência quanto a medida-F, consideradas nas avaliações, referem sempre os verdadeiros positivos presentes na lista a validar (para quem quiser considerar a abrangência unicamente como medida sobre o *corpus*, é suficiente assumir que a lista contém todos os verdadeiros positivos do *corpus*).

4.2 Validação por glossários especializados

Recuperamos um listado de termos do domínio geográfico do atlas geográfico do Brasil (IBGE 2010) e a taxonomia da ontologia de GeoNames traduzida no GoogleTranslate. O primeiro inclui o fator da concisão, assim os termos orientados são muito precisos, mas a lista é reduzida, e limitada portanto, na abrangência uma vez que não considera sinônimos nem termos geográficos não representados no mapeamento do IBGE. O segundo glossário maximiza a abrangência, porquanto inclui, por exemplo, termos relacionados com obras públicas e construções, mas fica mais limitado na precisão, visto que introduz ruído (falsos positivos) por desvios em escolhas léxicas da tradução automática (COSTA; DANIEL 2013). Criamos um terceiro glossário unindo os termos de IBGE e GeoNames com o objetivo de obter a máxima abrangência, suportando ainda um certo nível de ruído. Realizamos um primeiro teste de validação automática com estes três listados. A tabela 3 mostra os resultados. Para obtermos a máxima precisão, comprometemos a abrangência sobre os termos da lista a validar que, no melhor caso (validação com o glossário de 725 termos), apenas chega aos 65%.

Glossário	Precisão	Abrangência	Medida-F
IBGE	**100%**	54%	70%
GeoNames_trad	**100%**	42%	59%
IBGE ∪ GeoNames_trad	**100%**	65%	**79%**

Tabela 3: Comparação do efeito da validação da lista de candidatos a termos geográficos por meio de listados específicos do domínio geográfico.

Com o melhor resultado da validação por glossários criamos mais um listado especializado, nomeado com o código TR8IBGEGeonames (Tabela 4), composto por apenas 17 termos, com uma precisão do 100%.

***cidade**, reino, **ilha**, **porto**, **rio**, nome, lugar, fortaleza, terra, **costa**, dia, **barra**, **enseada**, ano, capitão, tempo, **vila**, filho, gente, **casa**, **estreito**, império, **mar**, partes, dom, estado, morte, naos, parte, povoação, armada, mercador, nao, **fazenda**, padre, senhor, **serra**, caminho, campo, embaixador, junco, **lago**, legoas, viagem, cousa, guerra, **monte**, partido, **ponta**, **castelo**.*

Tabela 4: Lista de termos candidatos extraídos do *subcorpus* do domínio geográfico (lista a validar), em negrito os termos avaliados automaticamente como verdadeiros positivos pelo glossário especializado de termos do domínio geográfico IBGE ∪ GeoNames_trad. Usamos a lista de termos avaliados como verdadeiros positivos (17 termos em negrito na tabela) como mais um novo listado do domínio que chamamos pelo código TR8IBGEGeonames.

4.3 Validação com uma base lexical-semântica difusa

Dada a limitação na abrangência dos resultados obtidos, desta vez, em vez de validarmos diretamente os candidatos, usamos as listas especializadas como semente, de onde procuramos termos relacionados numa base lexical, o CLIP 2.1. Realizamos uma série de experiências em que comparamos os resultados dos glossários aplicados no apartado anterior mais um quarto, TR8IBGEGeonames, obtido no melhor resultado da validação automática da lista (Tabela 4). O motivo para adicionarmos este quarto listado é avaliarmos também os efeitos da redução das pesquisas no CLIP 2.1 aos termos presentes no *corpus*. As experiências têm mais um outro objetivo para além de aumentarem a abrangência: achar os valores de associação semântica no CLIP 2.1 que ofereçam um melhor resultado na combinação da precisão, abrangência e medida-F.

O procedimento de pesquisa tem duas fases**,** seguidas de uma avaliação.

4.3.1 Pesquisa de synsets relacionados com o glossário

Recuperamos de CLIP 2.1 todos os *synsets* que contêm como mínimo um dos termos da lista semente, a partir de um valor de associação dado. Por exemplo, a lista semente TR8IBGEGeonames (termos em negrito na tabela 4) contém o termo *cidade*. Para a recuperação de *synsets* no CLIP 2.1, o algoritmo vai recuperar todos os *synsets* em que *cidade* apareça no CLIP2.1. Se, posteriormente, adicionamos um valor de associação, por ex. 0.5, vai recuperar apenas aqueles *synsets* em que o termo *cidade* apareça com um valor de associação igual ou acima dos 0.5.

4.3.2 Pesquisa dos termos da lista a validar dentro dos *synsets* selecionados

Dentro dos *synsets* recuperados pela lista semente, pesquisamos os termos da lista a validar a partir de um valor de associação dado. Se achamos o termo, consideramos que é do domínio geográfico; se não, fica fora como negativo. Por exemplo, com a lista TR8IBGEGeonames (termos em negrito na tabela 4) como semente, *capitão* não aparece em nenhum dos *synsets* recuperados, consequentemente, não é avaliado como pertencente ao domínio geográfico. Os termos *campo* e *castelo* também não aparecem na lista semente, não obstante, têm valores semânticos dentro dos *synsets* recuperados ao pesquisar no CLIP 2.1 e, portanto, são processados como pertencentes ao domínio geográfico. O resultado final é uma nova lista composta por aqueles termos que foram avaliados como pertencentes ao domínio.

cidade, *reino,* ***ilha***, ***porto***, ***rio***, *nome,* ***lugar***, ***fortaleza***, ***terra***, ***costa***, *dia,* ***barra***, ***enseada***, *ano, capitão, tempo,* ***vila***, *filho, gente,* ***casa***, ***estreito***, *império,* ***mar***, *partes, dom, estado, morte, naos, parte,* ***povoação***, *armada, mercador, nao,* ***fazenda***, *padre, senhor,* ***serra***, *caminho,* ***campo***, ***embaixador***, *junco,* ***lago***, *legoas, viagem, cousa, guerra,* ***monte***, *partido,* ***ponta***, ***castelo***.

Tabela 5: Lista de termos candidatos extraídos do *subcorpus* do domínio geográfico (lista a validar). Em negrito os termos avaliados como positivos ao pesquisar TR8IBGEGeonames com uma medida de associação de 0.15 para a recuperação de *synsets* e 0.2 para a extração de termos dentro dos *synsets* recuperados.

4.3.3 Avaliação

A lista obtida em cada teste é avaliada com base na lista de verdadeiros positivos totais (termos em negrito na tabela 2) que fora validada manualmente. Como exemplo, a tabela 5 mostra os resultados obtidos com o glossário TR8IBGEGeonames como semente. Os termos em negrito não sublinhados são verdadeiros positivos; os negritos sublinhados, falsos negativos. Ao avaliar os resultados respeitantes à lista de verdadeiros positivos validada manualmente (Tabela 2), obtemos uma precisão de 0.96, abrangência de 0.85 e medida-F de 0.9.

5. Resultados

As experiências permitiram avaliar o desempenho dos glossários e otimizar os valores de corte para a recuperação e pesquisa de *synsets* no CLIP 2.1. A figura 1 mostra os resultados obtidos em 400 rondas (4 glossários x 10 rondas de recuperação de *synsets* x 10 rondas de pesquisa de termos nos *synsets* recuperados). A lista mais específica do *corpus* (TR8IBGEGeonames) obtém os melhores resultados, com valores de 90% na medida-F. De salientar que a lista com menor número de termos (17) - portanto, com menor trabalho de pesquisa - atingiu os resultados mais altos. Não obstante, todas as listas especializadas conseguem, nas suas melhores configurações, resultados próximos.

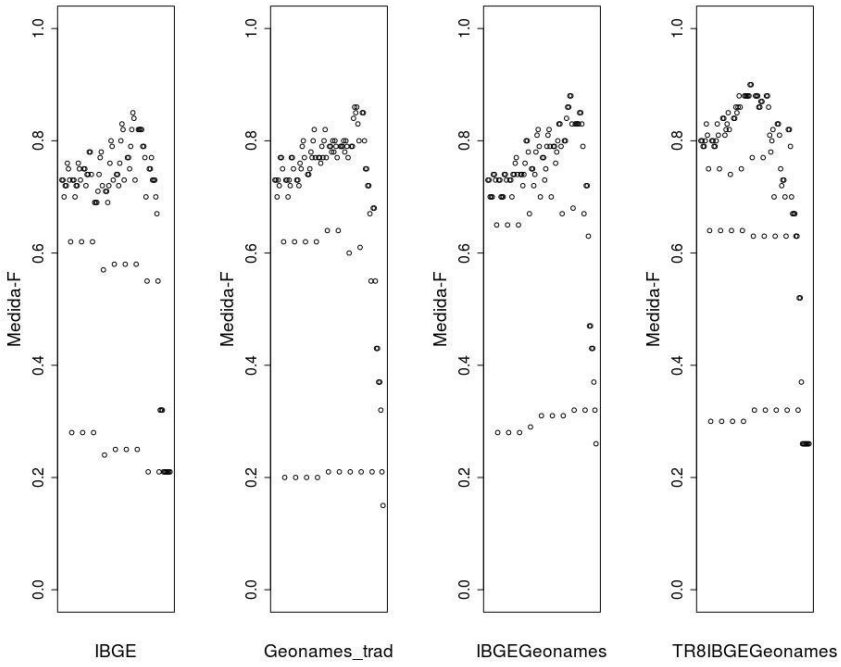

Fig.1: Resultado da validação por meio da base lexical segundo os glossários especializados usados como lista semente para a seleção de *synsets*.

No que diz respeito aos valores de corte na associação dos termos na base lexical, a tabela 6 mostra os melhores resultados para as três medidas consideradas. A abrangência aumenta pela redução do valor de corte. Achamos, assim, o melhor resultado da medida-F (84%), com um 100% de abrangência, com os valores de associação 0.05 na seleção e 0.01 para a limitação de pesquisa dentro do *synset*. Uma precisão do 100% mantém uma abrangência de 81%, com o valor de seleção de *synset* em 0.15 e recuperação de termos dentro do *synset* em 0.25. Os resultados mais baixos (Fig.1) são resultado da aplicação de medidas muito restritivas, que consideram termos com uma medida de corte ≥ 1.5.

Precisão	**Abrangência**	**Medida-F**	**Seleção de *synset***	**Seleção de termos no *synset***
100%	81%	90%	0.15	0.25
72%	100%	84%	0.05	0.01

Tabela 6: Melhores resultados de precisão e abrangência a respeito da medida-F e valores de corte na medida de associação semântica estabelecida pelo CLIP 2.1

6. Discussão

Os melhores resultados que achamos na literatura para a extração de termos combinam técnicas de PLN e estatísticas para obter listas de candidatos. Operando sobre uma lista obtida pela aplicação de técnicas convencionais, pesquisamos um método de melhoria que continue o processo de modo automático. Uma das soluções consideradas na literatura relacionada é a aplicação de listas especializadas. Consideramos dois glossários para as experiências: um que otimiza a precisão (IBGE) e outro que otimiza a abrangência (GeoNames_trad). Da sua reelaboração obtemos mais duas listas, uma da fusão de ambos (IBGE ∪ GeoNames_trad), outra da redução dos termos dos glossários àqueles que apenas achamos no *corpus* (TR8IBGEGeonames).

No que diz respeito à simples aplicação das listas especializadas, ao ampliar os termos contidos nos glossários com outros relacionados semanticamente (sinonímias difusas), a base lexical permitiu uma melhoria da abrangência até os 100% (Tabela 6), no melhor dos resultados das listas especializadas, limitada ao 65% (Tabela 3). É também possível atingir uma precisão de 100% e melhorar ainda a abrangência, no melhor resultado, até 81%, conseguindo um incremento de mais de 10 pontos na medida-F a respeito das listas especializadas (cuja precisão é também de 100%).

As experiências centraram-se em determinar os melhores valores para duas variáveis: o listado especializado (glossários) usado como semente para a recuperação e os valores de corte para a seleção de *synsets* e os termos neles incluídos. A respeito do glossário, aquele com menor número de termos (TR8IBGEGeonames com 17 termos - face a IBGE com 126 termos, GeoNames_trad com 667 e IBGE U GeoNames_trad com 725), porém o mais específico do corpus - é que obtém os melhores resultados, indicando que a proximidade da lista ao texto pode ser um fator relevante para o desempenho na recuperação de termos de domínio. A necessidade de considerar a variável do valor de associação semântica no CLIP 2.1 fica patente pelo fato de valores

muito restritivos (particularmente ≥1.5) oferecem resultados não desejados (os mais baixos na fig. 1). Como esperado, a redução do valor de corte aumenta a abrangência, no entanto, tanto para a seleção de *synsets* quanto para a posterior pesquisa dos termos neles incluídos, é possível estabelecer limites que reduzam o número de *synsets* e termos a avaliar, mantendo uma abrangência de 100%.

7. Conclusão

O uso de uma base lexical como ferramenta de validação dos candidatos obtidos pelas técnicas mais comuns de extração de termos permitiu a melhoria dos resultados através da simples aplicação de uma lista especializada em todas as métricas consideradas. O procedimento consistiu em selecionar um glossário geográfico e pesquisar os seus termos na base lexical CLIP 2.1 em que os termos aparecem agrupados em *synsets* com uma medida de associação semântica. Se achamos o termo pesquisado num *synset*, guardamos este grupo de termos. Se estabelecemos uma medida de corte, o procedimento é o mesmo mas só se guarda o *synset* em que o termo pesquisado aparece por cima do valor considerado. Para validar uma lista de termos geográficos que extraímos de um *corpus* do domínio geográfico, pesquisamos cada termo, primeiramente validado num glossário geográfico especializado, nos *synsets* do CLIP 2.1 e aplicamos uma nova medida de corte para a seleção dos termos dentro do *synset*. Deste modo melhoramos a abrangência dos listados especializados, ao aumentá-los com termos numa relação (quantificável) de proximidade semântica.

A vantagem da base lexical difusa não se limita à melhoria na abrangência. Consideramos que o seu aproveitamento é ainda maior, uma vez que permite configurações específicas aplicando valores de corte sobre a medida de associação semântica. Nas nossas experiências foi possível manter, com valores da medida-F > 80%, quer a precisão, quer a abrangência, no 100%. A base lexical CLIP 2.1 consegue estes resultados independentemente

do procedimento e métricas usados para chegar à lista de termos candidatos. Consideramos que a sua aplicação pode melhorar o presente estado da arte na extração de termos de domínio. O feito de ser de livre disposição faz que recomendemos a sua aplicação tanto em soluções práticas como em novos testes com configurações alternativas à por nós apresentada neste trabalho.

Referências bibliográficas

ALMEIDA, G. M. B, ALUÍSIO, S. M., & OLIVEIRA, L.H.M. O método em terminologia: revendo alguns procedimentos. In: Isquerdo, A. N.; Alves, I. M. *Ciências do léxico: lexicologia, lexicografia, terminologia*. Campo Grande/São Paulo: Editora da UFMS/Humanitas, 2007, vol. III, pp. 409-420. Disponível em: http://www.geterm.ufscar.br/textospublicados/o_metodo_em_terminologia_%20revendo_alguns_procedimentos.pdf

CONRADO, M. D. S. *Extração automática de termos simples baseada em aprendizado de máquina*. Universidade de São Paulo: Tese de doutoramento, 2014. Disponível em: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-11082014-103430/publico/TeseMerley_revisada.pdf.

CONRADO, M. S., FELIPPO, A., PARDO, T. A .S., & REZENDE, S. O. A survey of automatic term extraction for Brazilian Portuguese. *Journal of the Brazilian Computer Society*, vol. 20, n. 12, 2014, pp. 1-28. Disponivel em: https://www.researchgate.net/publication/265335491_A_survey_of_automatic_term_extraction_for_Brazilian_Portuguese

CONRADO, M. S., PARDO, T. A .S., & REZENDE, S. O. The main challenge of semi-automatic term extraction methods. In: *Proceedings of the 11st International Workshop on Natural Language Processing and Cognitive Science - NLPCS*. Venice, 2014, 27-29. Disponível em: http://conteudo.icmc.usp.br/pessoas/taspardo/NLPCS2014-ConradoEtAl.pdf

COSTA, G. C., & DANIEL, F. G. Google Tradutor: Análise de utilização e Desempenho da Ferramenta. *Tradterm*, vol. 22, 2013, pp. 327-361. Disponível em: http://www.revistas.usp.br/tradterm/article/view/69145/71600

FELIPPO, A., & ALMEIDA, G. M. B. Uma metodologia para o desenvolvimento de Wordnets terminológicas em português do Brasil. *Tradterm*, vol. 16, 2010, pp. 365-395. Disponível em: http://www.revistas.usp.br/tradterm/article/view/46325/50088

FEINERER, I., HORNIK, K., & MEYER, D. Text mining infrastructure in R. *Journal of statistical software*, vol. 25, n. 5, 1-54. Disponível em: https://www.jstatsoft.org/article/view/v025i05

FELLBAUM, C. Introduction. In: Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT, 1998, pp. 1-19.

GONÇALO OLIVEIRA, H., & GOMES, P. Onto. PT: automatic construction of a lexical ontology for Portuguese. In: *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. Lisbon, 2010, pp. 199-211. Disponível em: http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_Gomes_STAIRS2010.pdf

GONÇALO OLIVEIRA, H., & GOMES, P. ECO and Onto. PT: a flexible approach for creating a Portuguese wordnet automatically. *Language resources and evaluation*, vol. 48, n.2, 2014, pp. 373-393. Disponível em: http://link.springer.com/article/10.1007%2Fs10579-013-9249-9

GONÇALO OLIVEIRA, H., PAIVA, FREITAS, C., RADEMAKER, A., REAL, L., & SIMÕES, A. As wordnets do português. *Oslo Studies in Language*, vol. 7, n. 1, 2015, pp. 397-424. Disponível em: https://www.journals.uio.no/index.php/osla/article/view/1445/1342

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA [IBGE]. *Glossário dos Termos Genéricos dos Nomes Geográficos Utilizados no Mapeamento Sistemático do Brasil. Vol. 2*. Rio de Janeiro: Ministério do Planejamento, Orçamento e Gestão Instituto Brasileiro de Geografia e Estatística – IBGE, 2015. Disponível em: http://www.ibge.gov.br/home/geociencias/cartografia/glossario_termos_genericos_v2.shtm

LOPES, L., FERNANDES, P., & VIEIRA, R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. *Knowledge-Based Systems*, vol. 97, Abril 2016, pp. 237-249.

SALTON, G., & BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, vol. 24, n. 5, 1988, pp. 513-523.

SANTOS, F., & GONÇALO OLIVEIRA, H. Descoberta de Synsets Difusos com base na Redundância em vários Dicionários. *Linguamática*, vol. 7, n. 2, 2015, pp. 3-17.

SINGHAL, A., SALTON, G., & BUCKLEY, C. Length normalization in degraded text collections. In: *Fifth Annual Symposium on Document Analysis and Information Retrieval*, 1996, pp. 149-162. Disponível em: http://www.singhal.info/ocr-norm.pdf

TEIXEIRA, R. B. S. Análise do desempenho de extratores automáticos de candidatos a termos: proposta metodológica para tratamento de filtragem dos dados. *Tradterm*, vol. 18, 2011, pp. 297-319. Disponível em: http://www.revistas.usp.br/tradterm/article/view/36765/39487

WENDT, I. S., LOPES, L., MARTINS, D., VIEIRA, R., & LIMA, V. L. S. Geração automática de glossários de termos específicos de um corpus de Geologia. In: *3º ONTOBRAS. Seminário de Pesquisa em Ontologia no Brasil. 30 e 31 de*

Agosto de 2010. Florianópolis / SC. Florianópolis: Anais do 3º Seminário de Pesquisa em Ontologia no Brasil, 2010. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/ontobras/2010/0025.pdf

ZAPPAROLI, Z. M. Tratamento de corpora informatizados por programas de análise linguística para estudos do português falado de São Paulo. *Boletim da Academia Galega da Língua Portuguesa*, vol. 3, 2010, pp. 87-112.