Mapeamento de Candidatos

Alexandre Cançado Cardoso

Intrinsic Soluções em Informática M.E.

1 Introdução

Em um processo de anotação semântica a definição dos possíveis recursos candidatos para cada nome de entidade é uma tarefa fundamental. Pois é a partir destas relação que as anotações paras as entidades interessantes serão escolhidas. O conjunto das relações nome de entidades, recursos é denominado de mapa de candidatos. E pode ser definido de diversas formas.

Podendo o uso das diferentes formas de obtenção dos candidatos gerar mapas de qualidades distintas. Neste contexto, a avaliação de quais são as melhores estratégias de extração de candidatos se torna importante. Nas seções a seguir, quatro destas técnicas serão comparadas.

2 Estratégias de Extração de Mapa de Candidatos

As estratégias de extração de mapa de candidatos abordadas por este trabalho são a partir dos títulos, dos redirecionamentos, das desambiguações e das ocorrências no conteúdo das páginas.

O projeto DBpedia disponibilizam os arquivos com os títulos destas, redirecionamentos que levam para cada uma, e desambiguações (chamados, respectivamente, por: labels, redirects e disambiguations **??**). E é possível obter o dump do conteúdo todas as páginas em uma língua na Wikipédia **??**. A partir destes o DBpedia Spotlight consegue extrair mapeamentos de candidatos segundo as quatro estratégias descritas a seguir respectivamente. Na quais o recurso "Berlin" da DBpedia em inglês (http://dbpedia.org/resource/Berlin) será utilizada como exemplo.

2.1 Extração a partir dos títulos

Toda a página na DBpedia que descreve um recurso é identificada por uma URL composta do endereço da DBpedia para o determinado idioma de um identificador de que se trata de um recurso, idêntico para todas as identidades, e, por último, de um rótulo (*label*) exclusivo da entidade, o qual também é chamado de título. Do exemplo teríamos:

- Parte comum: "http://dbpedia.org/page/"
- Rótulo/Título: "Berlin"

A estratégia de extração de candidatos a partir dos títulos consiste em considerar o rótulo de todos os recursos como uma *surface forms* da qual o respectivo recurso é candidato. Portanto, um candidato para a *surface form*, "berlin" footnoteNeste relatório cosideramos as variações de maíuscula e minúsculas, como sendo a mesma *surface form*. Ou seja, "berlin" é considerado equivalente a: "Berlin", "BERLIN", etc. é o resurso "http://dbpedia.org/resource/Berlin". Por esta estratégia tem-se, a natural relação de que todo o recurso da DBpedia no idioma será candidato de pelo menos a *surface form* que o entitula.

2.2 Extração a partir dos redirecionamentos

Uma recurso na DBpedia e na Wikipédia pode ser alcançado não apenas por seu título, podendo, diversos, serem acessados por outras formas de referenciamento. Sendo estes identificadores secundário chamados de redirecionamentos. Por exemplo, o recurso "http://dbpedia.org/resource/Berlin" também pode ser acessada pelos redirecionamentos relacionado pelo elemento "DBpedia-owl:wikiPageRedirects" da ontologia da DBpedia (descritos em http://dbpedia.org/resource/Berlin). A lista abaixo apresenta alguns destes:

- Berlin,_Germany
- City_of_Berlin
- Capital_of_East_Germany
- Berlin-Zentrum
- Berlin.de
- Berlin_(Germany)
- Federal_State_of_Berlin
- Historical_sites_in_berlin
- Land_Berlin

É importante ressaltar que em qualquer uma das estratégia de extração de candidato o DBpedia Spotlight desconsidera durante os caracteres especiais adicionados pela DBpedia, afim de obter as *surface forms* no formato original em que aparecem em textos. Obtendo, assim, a lista a seguir:

- "berlin, germany"
- "city of berlin"
- "capital of east germany"
- "berlin-zentrum"
- "berlin.de"
- "berlin (germany)"
- "federal state of berlin"
- "historical sites in berlin"
- "land berlin"

Observa-se que da mesma forma que uma ocorrência da palavra "berlin" em um texto tem chanses de se referir ao recurso "http://dbpedia.org/resource/Berlin", isto é válido para "city of berlin", "berlin, germany" e as demais obtidas do redirecionamento. Por isto, a estratégia de extração de candidatos a partir dos redirecionamentos, considerará as lista acima como *surface forms* para as quais o recurso "http://dbpedia.org/resource/Berlin" é um candidato.

2.3 Extração a partir das desambiguações

Uma expressão pode ter diversos significados, estas ambiguidades são relacionadas nas páginas de desambiguação da DBpedia (e Wikipédia). O DBpedia Spotlight é capaz de extrair utilizando-se de um processo inverso ao realizado pelo estratégia a partir dos redirecionamentos.

Seja a página de desambiguação da DBpedia: "http://dbpedia.org/page/Berlin_(disambiguation)". Removendo as estruturas utilizadas pela DBpedia para construção da URL e identificação de que se trata de uma página de desambiguação, tem-se a *surface form*: "berlin"
footnoteE realizando as mesmas formatações descritas na Seção 2.1.

As páginas de desambigução apresentam no elemento "dbpedia-owl:wikiPageDisambiguates" da ontologia todos os recursos relacionados com a *surface form* obtida do título da página de desambiguação. Alguns destes recursos para a página de desambiguação de exemplo estão listados abaixo:

- http://dbpedia.org/resource/Berlin
- http://dbpedia.org/resource/Berlin,_New_Hampshire
- http://dbpedia.org/resource/Berlin,_Wisconsin
- http://dbpedia.org/resource/Berlin,_Marathon_County,_Wisconsin
- http://dbpedia.org/resource/Berlin_(Amtrak_station)
- http://dbpedia.org/resource/Berlin_(band)
- http://dbpedia.org/resource/Mount_Berlin
- http://dbpedia.org/resource/SS_Berlin
- http://dbpedia.org/resource/Berlin_(comic)
- http://dbpedia.org/resource/Berlin_(surname)
- http://dbpedia.org/resource/East_Berlin_(disambiguation)
- http://dbpedia.org/resource/West_Berlin_(disambiguation)

Observa-se que além do recurso utilizado como exemplo para as estratégias anteriores (Seções 2.1 e 2.2), o qual é a entidade referente a capital da Alemanha, a expressão "berlin" também está associada a outras cidades, tal qual a entidades de outras natureza como: estação de trem, banda de música, monte, embarcação, série de história em quadrinhos, sobrenome.

Portanto, está estratégia de extração de mapa de candidatos irá considerar todos estes recursos como candidatos para a *surface form* extraída do título da página de desambiguação, no caso "berlin". Esta, ainda, terá como candidatos os recursos relacionados as páginas de desambiguação que estiverem listadas no elemento "dbpedia-owl:wikiPageDisambiguates". Para o exemplo, todos os recursos obtidos pela estratégia de desambiguação para as páginas "http://dbpedia.org/resource/East_Berlin_(disambiguation)" e "http://dbpedia.org/resource/West_Berlin_
serão candidatos para "berlin", tal qual para a *surface form* obtida do título de suas respectivas páginas.

2.4 Extração a partir das ocorrências

Diferentemente das estratégias anteriores, a extração a partir das ocorrências no contúdo de uma página da Wikipédia não utiliza-se da estrutura do nome

dos recursos ou dos elementos da ontologia definida na DBpedia. Por sua vez, é realizado processando o conteúdo cada página da Wikipédia (as quais são obtidas através de dump desta para um idioma realizado pelo software livre **??**).

Este processamento é feito de forma a identificar as expressões de mais relevantes no texto contido na página. As quais serão consideradas *surface forms* relacionadas ao recurso da página. Por exemplo, seja a página da Wikipédia: "http://en.wikipedia.org/wiki/Berlin" e algumas expressões relevantes identificadas listadas a seguir:

- "berlin"
- "germany's largest city"
- "capital city of germany"
- "capital of the kingdom of prussia"
- "berliner"

Sabe-se que que o recurso da DBpedia referente a está página é o recurso: "http://dbpedia.org/resource/Berlin". O qual será mapeado como candidato para todas as *surface forms* listadas acima.

3 Experimentos

Como padrão de comparação (*Gold Standard - GS*) para a avaliação da qualidade destes quatro mapas, foram utilizados três cospora obtidos na literatura com textos em inglês: Milne & Witten Corpus ([**?**]), CSAW Corpus ([**?**]) e Aida-CoNLL-Yago2 Corpus ([**?**]). Todos estes compostos A Tabela **??** apresenta suas principais características:

Table 1: Características dos corpora utilizados

	M&W	**CSAW**	**Aida CoNLL**
Estilo dos Artigos	Jornalístico	Jornalístico e *Wikipédia*	Jornalístico
Número de Artigos	50	103	1393
Número de Anotações	706	12099	34929
Número de *Surface Forms* Distintas	579	11015	4847

Inicialmente foram construídos os mapas de candidatos (*Candidate Maps - CM*) por cada uma das estratégias descritas na Seção 2 separadamente. Obtendo, assim, os mapas a partir: dos títulos (T), dos redirecionamentos (R), das desambiguações (D) e das ocorrências (O). Isto foi realizado utilizando o Algoritmo **??** tendo como entrada o *dump* de todos os artigos da Wikipédia **??** em inglês (através do software **??**) e os arquivos labels.nt, redirects.nt e disambiguations.nt obtidos do projeto DBpedia para o inglês **??**. A Tabela **??**

Analisando a Tabela **??** e a última linha da Tabela **??** observa-se que existe muito mais *surface forms* distintas nos mapas de candidatos do que nos

Table 2: Número de *Surface Forms* Distintas nos Mapa de Candidatos

Estratégia de Extração	M&W	CSAW	Aida CoNLL
Títulos	2490150	2490150	2490150
Redirecionamentos	3285816	3285816	3285816
Desambiguações	179447	179447	179447
Ocorrências	2413762	2413762	2413762

gold standards. Portanto, foi verificado quantas destas estão em ambos desconsiderando as repetições, isto é apresentado pela Tabela **??**, enquanto a Tabela **??** mostra a quantidade de **??** distintas contidas no *gold standard* mas ausentes nos mapas de candidatos.

Table 3: Número de *Surface Forms* Distintas na Interseção dos GS e CM

Estratégia de Extração	M&W	CSAW	Aida CoNLL
Títulos	198	1039	4659
Redirecionamentos	191	1009	3617
Desambiguações	132	615	2444
Ocorrências	71	369	2476

Table 4: Número de *Surface Forms* Distintas dos GS Não Contidas nos CM

Estratégias de Extração	M&W	CSAW	Aida CoNLL
Títulos	381	3808	6356
Redirecionamentos	388	3838	7398
Desambiguações	447	4232	8571
Ocorrências	508	4478	8539

Para cada anotação em um *gold standard* foi verificado se o recurso para o qual estava anotada era um dos candidatos relacionados a *surface form* pelo respectivo mapa. As Tabelas **??** e **??** apresentam as quantidades de avaliações positivas e negativas, respectivamente:

Note que as Tabelas **??** e **??** contém valores maiores que os da Tabela **??** em algumas células. Isto se deve ao fato de que na Tab. **??** não contar as repetições de *surface forms*, e as demais avaliar estas repetições. O que é necessario pois duas ocorrências da mesma *surface form* podem ter sido anotadas para diferentes recursos, onde estes, por sua vez, podem, idependentemente um do outro, estar ou não contidos no mapa de candidatos. A Tabela **??** apresenta o número de *surface forms* contidas ambas no *gold standard* e nos mapas de candidatos, contando inclusive as repetições (isto é o mesmo que a soma das células das Tabelas **??** e **??**):

Table 5: Número de Recursos Anotados no GS Candidatos da Respectiva *Surface Forms* no CM

Estratégias de Extração	**M&W**	**CSAW**	**Aida CoNLL**
Títulos	194	1606	10260
Redirecionamentos	113	928	4070
Desambiguações	39	281	1727
Ocorrências	11	182	2008

Table 6: Número de Recursos Anotados no GS que Não são Candidatos da Respectiva *Surface Forms* no CM

Estratégias de Extração	**M&W**	**CSAW**	**Aida CoNLL**
Títulos	61	780	9084
Redirecionamentos	154	1459	15119
Desambiguações	146	1361	13543
Ocorrências	73	436	2969

Table 7: Número de *Surface Forms* Contando Repetições na Interseção dos GS e CM

Estratégias de Extração	**M&W**	**CSAW**	**Aida CoNLL**
Títulos	255	2386	19344
Redirecionamentos	267	2387	19189
Desambiguações	185	1642	15270
Ocorrências	84	618	4977

Por último, foram obtidas duas métricas de sensibilidades (baseado em [**?**]):

- *Surface Forms Recall* (Tabela **??**) - Razão da quantidade de *surface forms* do GS no CM pela quantidade de *surface forms* distintas de GS;
- *Resources Recall* (Tabela **??**) - Razão dos recursos recuperados (Tabela **??**) pelas *surface forms* recuperadas (Tabela **??**).

Table 8: Surface Forms Recall

Estratégias de Extração	**M&W**	**CSAW**	**Aida CoNLL**
Títulos	0.3419689119	0.2143593976	0.4229686791
Redirecionamentos	0.3298791019	0.2081700021	0.328370404
Desambiguações	0.2279792746	0.1268826078	0.221879256
Ocorrências	0.1226252159	0.0761295647	0.224784385

Table 9: Resources Recall

Estratégias de Extração	**M&W**	**CSAW**	**Aida CoNLL**
Títulos	0.9797979798	1.5457170356	2.202189311
Redirecionamentos	0.5916230366	0.9197224975	1.1252419132
Desambiguações	0.2954545455	0.4569105691	0.706628478
Ocorrências	0.1549295775	0.4932249322	0.81098546

4 Conclusão

A **TODO**: conclusão