Mapeamento de Candidatos

Alexandre Cançado Cardoso

Intrinsic Soluções em Informática M.E.

Abstract.

1 Introdução

Em um proceso de anotação semântica a definição dos possiveis 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 seram 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 titulos, dos redirecionamentos, das disambiguações e das ocorrências no conteúdo das páginas.

2.1 Extração a partir dos títulos

A **TODO**: seção

2.2 Extração a partir dos redirecionamentos

A **TODO**: seção

2.3 Extração a partir das disambiguações

A **TODO**: seção

2.4 Extração a partir das ocorrências

A **TODO**: seção

3 Experimentos

Inicialmente foram construidos os mapas de candidatos por cada uma das estratégias descritas na Seção 2 separadamente. Obtendo, assim, os mapas a partir: dos titulos (T), dos redirecionamentos (R), das disambiguações (D) e das ocorrências (O). Isto foi realizado utilizando o Algoritmo **??** tendo como entrada o *dump* de todos os arquivos da DBpedia **??** em inglês.

Para a avaliação da qualidade destes quatro mapas, foram utilizados três cospora obtidos na literatura com textos em inglês (Seção 3.1). Onde, para cada entidade destes foram verificadas se o recurso ao qual estavam anotados era um dos candidatos relacionados a entidade pelo mapa em questão (Algoritmo **??**). Então, foram calculadas as seguintes métricas (conforme proposto por [**?**]):

- *Acurácia* - é ... **TODO**: descrever
- **TODO**: descrição métricas utilizadas segundo Hachey

3.1 Corpora

Os corpora utilizados foram: Mille & 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**
Número de Artigos	50	103	1393
Número de Entidades	706	12099	34929
Estilo dos Artigos	Jornalístico	Jornalístico e *Wikipedia*	Jornalístico

3.2 Resultados

A **TODO**: Carolina

4 Conclusão

A **TODO**: Depende dos resultados