O paradigma de supervisão distante para extração de relações

Leonardo Martinez Ikeda
Orientador: Prof. Dr. Denis Deratani Mauá

Abordagens estudadas

Foco para dois dos principais trabalhos da área:

Mintz *et. al* (2009)

- Introdução do paradigma de supervisão distante
- Modelagem simples

Zeng *et. al* (2015)

- Modelagem simples
- Baseada em redes neurais, estado da arte da tarefa atualmente

+ outros trabalhos posteriores a cada um destes, tratando de suas principais deficiências

1 Extração de relações

O que é a extração de relações?

- Tarefa do processamento de linguagem natural
- **Objetivo:** extrair de um *corpus* de textos não estruturados **relacionamentos** expressos semanticamente entre **entidades**

"**Bill Gates** é um dos cofundadores da **Microsoft**"

→ (*cofundador*, *Bill Gates*, *Microsoft*)

"**Barack Obama** é casado com **Michelle Obama** desde 1992"

→ (*cônjuge*, *Barack Obama*, *Michelle Obama*)

O que é a extração de relações?

- **Entidades:** objetos que possam ser unicamente identificados (pessoas, lugares, eventos, etc.)
- **Relacionamento:** associação entre entidades
- Formalmente, uma **relação** é uma tripla **(r, e1, e2)**, denotando o relacionamento do tipo r entre as entidades

Abordagens usuais

Aprendizado supervisionado

Extração de *features* léxicas, semânticas e sintáticas de um conjunto de textos manualmente anotados com as relações expressas

Aprendizado não supervisionado

Extração e *clusterização* de sequências de palavras entre entidades de textos

Aprendizado semi-supervisionado

Extração a partir de exemplos sementes em um processo de aprendizado iterativo

(sementes > exemplos > padrões > novas relações > exemplos > ...)

Abordagens usuais

Aprendizado supervisionado

\+ Resultados obtidos apresentam alta precisão

– Rotular manualmente os dados é um processo lento e custoso

Aprendizado não supervisionado

\+ Facilmente aplicável para um grande volume de dados

– Resultados em formato pouco estruturado

Aprendizado semi-supervisionado

\+ Facilmente aplicável para um grande volume de dados

– Resultados de baixa precisão

2 Supervisão distante

Supervisão distante

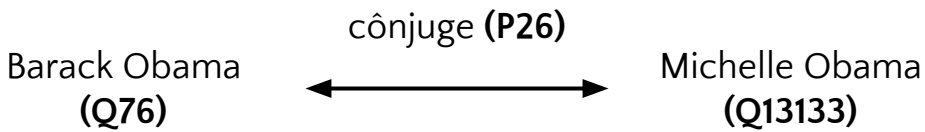
- Idéia central: rotular dados automaticamente explorando uma **base de conhecimento** externa

- Objetivo de unir os benefícios das diferentes abordagens
  - Utilizar grande volume de dados
  - Obter resultados com bom desempenho e formato

Supervisão distante

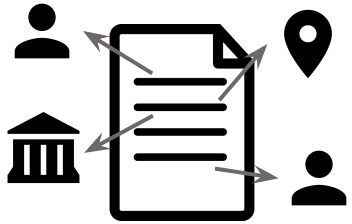
- **Bases de conhecimento:** bases que armazenam informações sobre entidades de forma estruturada
- Diversas bases de conhecimento disponíveis de forma aberta atualmente (Wikidata, YAGO, etc.)

Exemplo de informações extraídas do Wikidata:



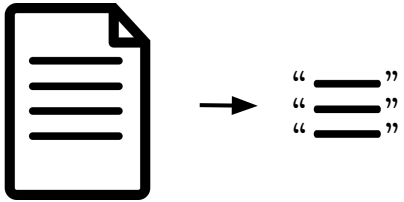
Supervisão distante

Corpus de textos



Reconhecimento de entidades

Ligação com a base de conhecimento



Segmentação em sentenças

Alinhamento e rotulação de sentenças

Supervisão distante

Alinhamento de sentenças com informações da base de conhecimento é baseado em uma **heurística:**

Se duas entidades estão relacionadas na base de conhecimento, então **todas** as sentenças contendo ambas as entidades expressam tal relacionamento

Supervisão distante

- **Exemplo:**

Sentença → Rótulo

"**Bill Gates** é um dos cofundadores da **Microsoft**" → **(cofundador)**

Base de conhecimento

Supervisão distante

Geração de um grande volume de dados para alimentar modelos de aprendizado

- **ACE**: **17.000** instâncias de **24** tipos de relações
- **Mintz**: **1.800.000** instâncias de **102** tipos de relações

Supervisão distante

Elaboração de *features* mais exatas, precisas

- Alta precisão, baixa revocação
  - Palavras entre as entidades
  - Janela de *k* palavras antes/após entidades
  - Etiquetas morfossintáticas de cada palavra
  - Caminho de dependência entre as entidades
  - ...

Supervisão distante

Extração de relações em nível de *bags* de sentenças:

- Vetor de *features* da sentença: conjunção de cada uma das *features* léxicas e sintáticas anteriores
- Vetor de *features* do *bag:* combinação dos vetores de *features* de cada sentença

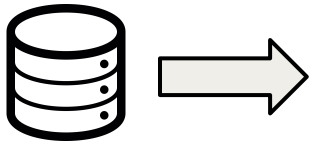
Supervisão distante

Principal problema:

Rótulos potencialmente ruidosos

Sentença

"**Bill Gates** deixou hoje o conselho da **Microsoft**" → Rótulo: **(cofundador)** **falso positivo**



Base de conhecimento

Bill Gates —*cofundador*→ **Microsoft**

Supervisão distante

Adoção de uma **heurística relaxada:**

Se duas entidades estão relacionadas na base de conhecimento, então **ao menos uma** das sentenças contendo ambas as entidades expressa tal relacionamento

Supervisão distante

Reduzir impacto de falsos positivos:

- Heurística relaxada
- Aprendizado múltipla instância
  - Rótulos em relação a *bags* de instâncias
  - *Bags* com **ao menos uma** instância positiva = rótulo positivo
- Aprendizado múltipla instância com múltiplos rótulos

3

Redes neurais convolucionais

Redes neurais convolucionais

- Modelagens baseadas em redes neurais convolucionais por partes
  - Redes neurais convolucionais para processamento de linguagem natural (*word embeddings* + convolução 1D)
  - *Embeddings* de posição: distância da palavra às entidades
  - Por partes: camada de *max pooling* por partes

    “A multinacional **Disney** comprou a **Pixar** em 2006”

    R1 R2 R3
- **Objetivo:** viabilidade de redes neurais + não depender da elaboração de *features*

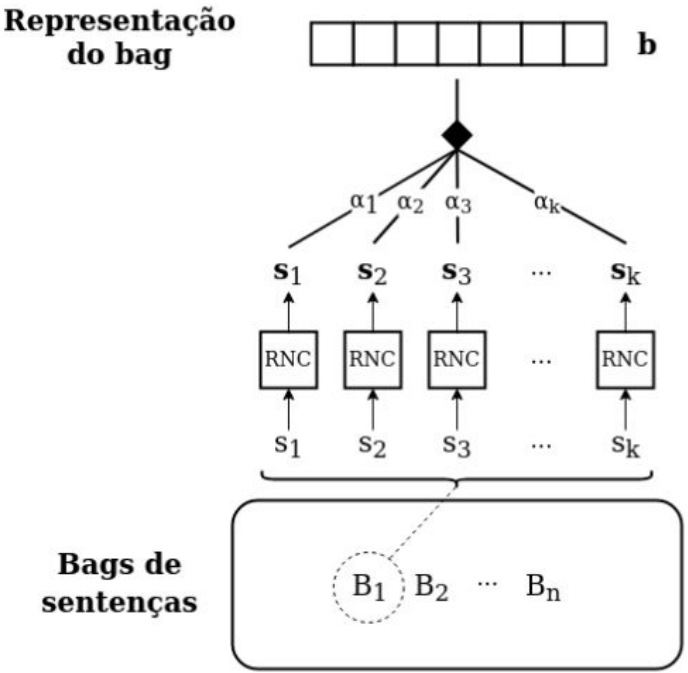
Redes neurais convolucionais

Tratamento de rótulos ruidosos:

- **Zeng *et al.* (2015):** Redução de impacto de ruídos somente considera sentença de maior *score* de um *bag*

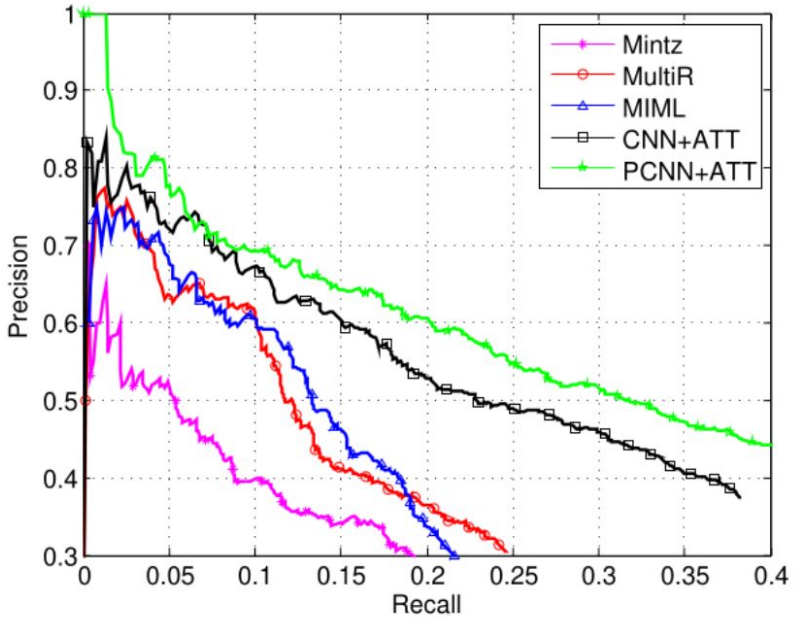
- **Y. Lin *et al.* (2016):** Mecanismo de atenção seletiva
  - Ajuste de pesos de cada sentença do *bag*
  - Representação vetorial do *bag*

Redes neurais convolucionais



Representação vetorial do *bag* substitui representação vetorial da melhor sentença

Resultados dos trabalhos



Fonte: Y. Lin *et al.* (2016) "Neural relation extraction with selective attention over instances".

PCNN obteve resultados próximos do MultiR ou MIML

Obrigado!

Dúvidas?

Página do TCC

- http://www.linux.ime.usp.br/~ikedaleo/mac0499