Processamento de Linguagem Natural

Aula 05

Extração de Informação
e Reconhecimento de Entidades

PUC Minas

Extração de Informação

- Explosão de informações na forma de notícias, arquivos corporativos, registros médicos, documentos governamentais, redes sociais, etc.
- Informações desestruturadas que dificultam interpretação

Extração de Informação

- Entradas dos sistemas de Extração de Informação são coleções de documentos
  - Emails, sites, artigos de jornal, relatórios, artigos científicos, pesquisas, blogs.
- Saída é uma representação da informação relevante a partir do documento de entrada, de acordo com algum critério específico.

Tarefas da Extração de Informação

- Extração de Informação faz parte dos estágios iniciais de várias tarefas de alto nível como Sistemas de Respostas Automáticas e Máquinas de Traduções.
- A extração de informação é útil para várias aplicações comerciais como BI e gestão do conhecimento.

Tarefas da Extração de Informação

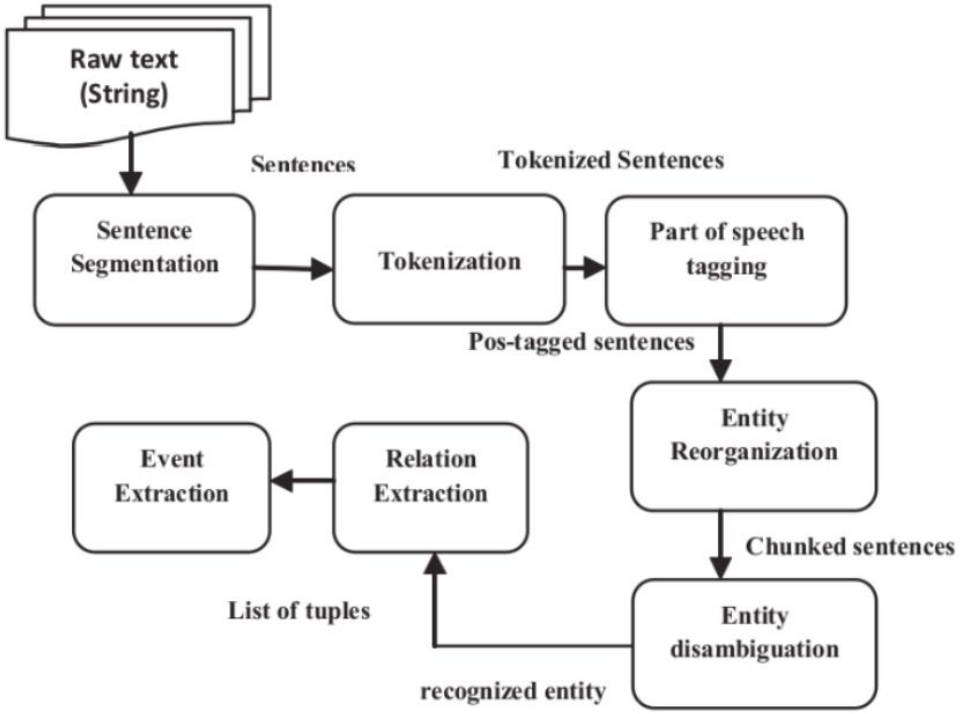


Fig.1 General Information Extraction Architecture. Adapted from (Costantino et al., 1997)

PUC Minas

Tarefas da Extração de Informação

- Várias sub-tarefas são envolvidas no processo da extração da informação, tais como:
  - Reconhecimento de Entidades Nomeadas
  - Relacionamento de Entidades Nomeadas
  - Extração de Informação Temporal
  - Construção de Base de Conhecimento

Reconhecimento de Entidades Nomeadas

- Named Entity Recognition (NER)
- Sistemas que precisam reconhecer entidades nomeadas como Pessoa(PER), Organização(ORG), Localização (LOC)

”Michael Jordan lives in United States”

Sistema NER extrai Michael Jordan como Pessoa e United States como nome de país, Localização.

PUC Minas

Relacionamento de Entidades Nomeadas

- Named Entity Linking (NEL) , Named Entity Disambiguation (NED) ou Named Entity Normalization (NEN)
- Tarefa de identificação da entidade correspondente a uma ocorrência particular no documento
- A referência a entidades em textos naturais pode ser ambígua.
    - Paris - Pode ser LOC (Capital da França) ou PER (Atriz Paris Hilton)

Extração de Informação Temporal

- Temporal information extraction ou event extraction
- Identificação de eventos - informação que pode ser ordenada em ordem temporal
- Expressão temporal se refere à detecção de frases na linguagem natural que denotem uma entidade temporal, intervalo, tempo/hora, frequência.

Extração de Informação Temporal

”President Barack Obama **yesterday** addressed the issue of nuclear deals at White House”.

Yesterday é o substantivo que se refere à informação temporal

Informações temporais são importantes quando queremos extrair informações estruturadas de textos em linguagem natural com algum critério temporal:

- Notícias
- Biografias
- Eventos

PUC Minas

Construção de Base de Conhecimento

- Bases de conhecimento (Knowledge Base) são utilizadas atualmente em várias aplicações diferentes
  - Sites de busca
  - Sistemas de apoio de decisão
  - Perguntas e respostas
  - Assistentes digitais: Siri, Cortana, Google Assistant
- São grandes em tamanho
  - Milhões de fatos, bilhões de entidades, milhões de relações
  - Ainda geram erro pela ausência de muitos fatos

Named Entity Recognition

- Todos os textos contém termos particulares que representam entidades específicas em seu contexto único.
  - Essas entidades são conhecidas como **Entidades Nomeadas.**
  - São termos que representam objetos do mundo real como pessoas, lugares, organizações.

Named Entity Recognition

- Como as entidades nomeadas geralmente são representadas por nomes próprios, uma abordagem *ingênua* seria encontrar todos os substantivos dentro do texto.

Named Entity Recognition

- **Named entity recognition** (NER) , também conhecida como **entity chunking/extraction** , é a técnica utilizada na extração de informação para:
  - Identificar entidades
  - Segmentar entidades
  - Classificar e categorizar dentro das classes pré-definidas

Named Entity Recognition

- Desafio de criar um algoritmo para aprender a identificar corretamente o que chamamos de entidades.

“O Rato roeu a roupa do rei de Roma”

- As entidades da frase acima seriam **Rato**, **Roma** e talvez **roupa**.

PUC Minas

Named Entity Recognition

Identificação

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Classificação

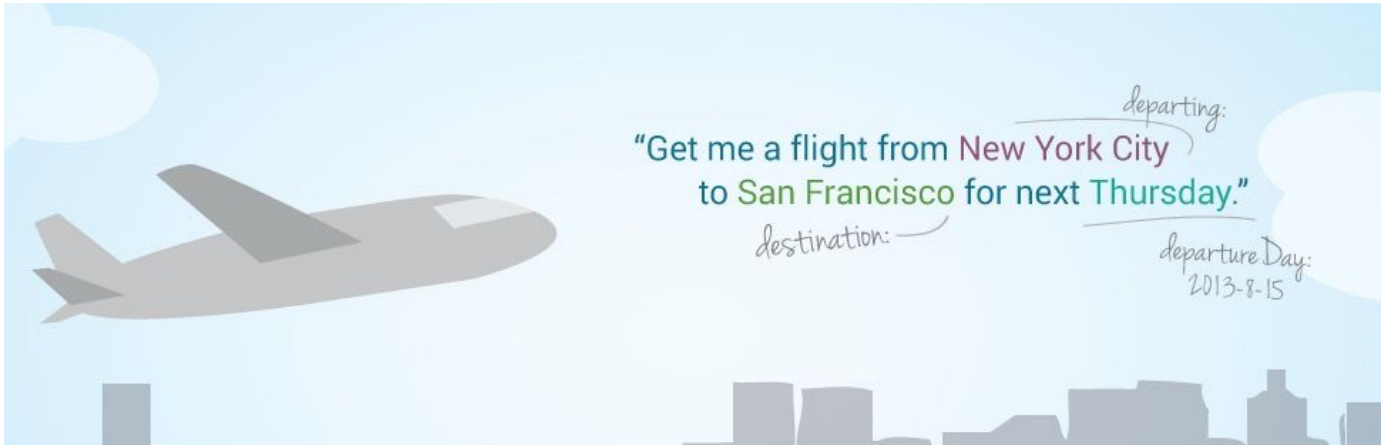
Pessoa
Data
Organização

Tipos Comuns de Entidades Nomeadas

NE Type	Examples
ORGANIZATION	*Georgia-Pacific Corp., WHO*
PERSON	*Eddy Bonte, President Obama*
LOCATION	*Murray River, Mount Everest*
DATE	*June, 2008-06-29*
TIME	*two fifty a m, 1:30 p.m.*
MONEY	*175 million Canadian Dollars, GBP 10.40*
PERCENT	*twenty pct, 18.75 %*
FACILITY	*Washington Monument, Stonehenge*
GPE	*South East Asia, Midlothian*

Para que serve o reconhecimento de entidades?

- Útil para criar um sistema que precise identificar o que um ser humano está falando ou digitando em um chat, por exemplo.
- Se o algoritmo consegue aprender as entidades nos dados, pode disparar ações específicas



PUC Minas

Aplicações do Reconhecimento de Entidades Reconhecidas

- Resumo de Grandes Documentos
  - Um sistema que utiliza RER poderia aprender a resumir textos grandes em pequenos a partir do conhecimento de entidades relevantes.

Aplicações do Reconhecimento de Entidades Reconhecidas

- Categorização de Reviews
  - Ao aplicar RER em comentários de usuários em portais, pode se gerar categorias que resumem todo o texto. Um exemplo é a Amazon que resume um review sobre um produto no portal.

Aplicações do Reconhecimento de Entidades Reconhecidas

- Chatbot de Auto-atendimento
  - Diversos Chatbots estão sendo construídos nos quais a base para seu funcionamento é a técnica de RER. Com o reconhecimento das entidades os robots tomam ações propícias conforme o usuário fala ou digita.

Aplicações do Reconhecimento de Entidades Reconhecidas

- Identificação e agendamento de Compromissos
  - A imagem de destaque desse artigo mostra um exemplo dessa aplicação. Usando RER um sistema consegue aprender o que é data, horário e uma cidade em um e-mail por exemplo. Com isso nas “mãos” o sistema poderia agendar um compromisso automaticamente na sua agenda.

Biblioteca SpaCy

- Biblioteca em Python para PLN em escala industrial.
- Desenvolvida pela Explosion AI, especificamente para uso em produção e ajudar a criar aplicações que conseguem processar e “entender” um grande volume de texto.
- Pode ser usada para extrair informações ou entendimento de linguagem natural ou pré-processar texto para deep learning.

https://spacy.io

PUC Minas

Biblioteca SpaCy

Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.

Biblioteca SpaCy

Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

Biblioteca SpaCy

Entre as várias funcionalidades, um ponto importante para nós é que a biblioteca provê o português como uma das línguas que trata:

```
import spacy
nlp = spacy.load('pt_core_news_sm')
```

PUC Minas