Extração de Informação (EI)

Jademir de Moura Barbosa Filho
Jorge Tiago Moura Cruz
Sara Inés Rizo Rodríguez
Teresa Rachel Pernambuco

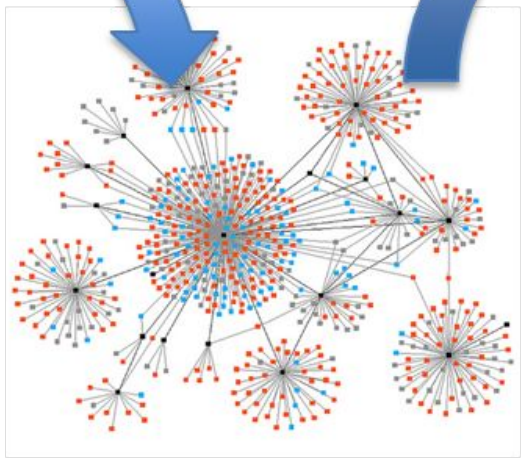
UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

ROTEIRO:

- INTRODUÇÃO
- ORIGEM
- TIPOS DE TEXTO
- TAREFAS EM EI
- TIPOS DE SISTEMAS
- EI EM MÍDIA SOCIAL
- EXTRAÇÃO DE INFORMAÇÃO ABERTA
- FERRAMENTAS
- REFERÊNCIAS

INTRODUÇÃO

Publications



Relationships

Structured Information

INTRODUÇÃO:

- A descoberta de conhecimento em bases de dados inicia-se com a seleção de um conjunto ou amostra de dados com os quais o processo de descoberta será realizado.

- Para isso utilizam-se técnicas de mineração que procuram por padrões e regularidades, por fim, as informações descobertas são interpretadas e avaliadas de forma que se selecione os conhecimentos úteis resultantes de todo este processo.

INTRODUÇÃO:

- Segundo Jim Cowie, sistemas de RI podem ser vistos como “colheitadeiras” que devolvem material útil de um vasto campo de materiais brutos. Com grande quantidades de informações potencialmente úteis em mãos, um sistema de EI pode, então, transformar o material bruto refinando e reduzindo-o à idéia do texto original.

EXEMPLO:

- *“Three bombs have exploded in north-eastern Nigeria, killing 25 people and wounding 12 in an attack carried out by an Islamic sect. Authorities said the bombs exploded on Sunday afternoon in the city of Maiduguri.”*

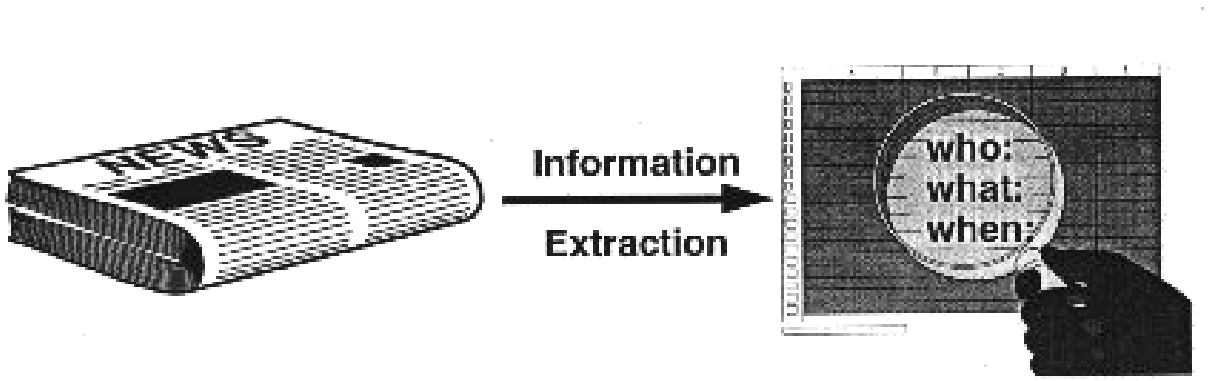
TYPE:	Crisis
SUBTYPE:	Bombing
LOCATION:	Maiduguri
DEAD-COUNT:	25
INJURED-COUNT:	12
PERPETRATOR:	Islamic sect
WEAPONS:	bomb
TIME:	Sunday afternoon

Extração de Informação vs. Recuperação de Informação (RI)

Recuperação de Informação

- Selecionar um subconjunto de documentos de uma coleção de documentos textuais relevantes para a consulta.
- Retorna uma lista de documentos ranqueados.

Extração de Informação vs. Recuperação de Informação (RI)



Extração de informação

- Extrair fatos dos documentos sobre seu conteúdo semântico.

HISTÓRIA:

- JASPER (1980s)
  - Um dos primeiros Sistemas do âmbito comercial a usar EI.
  - Sistema sobre notícias financeiras em tempo real.

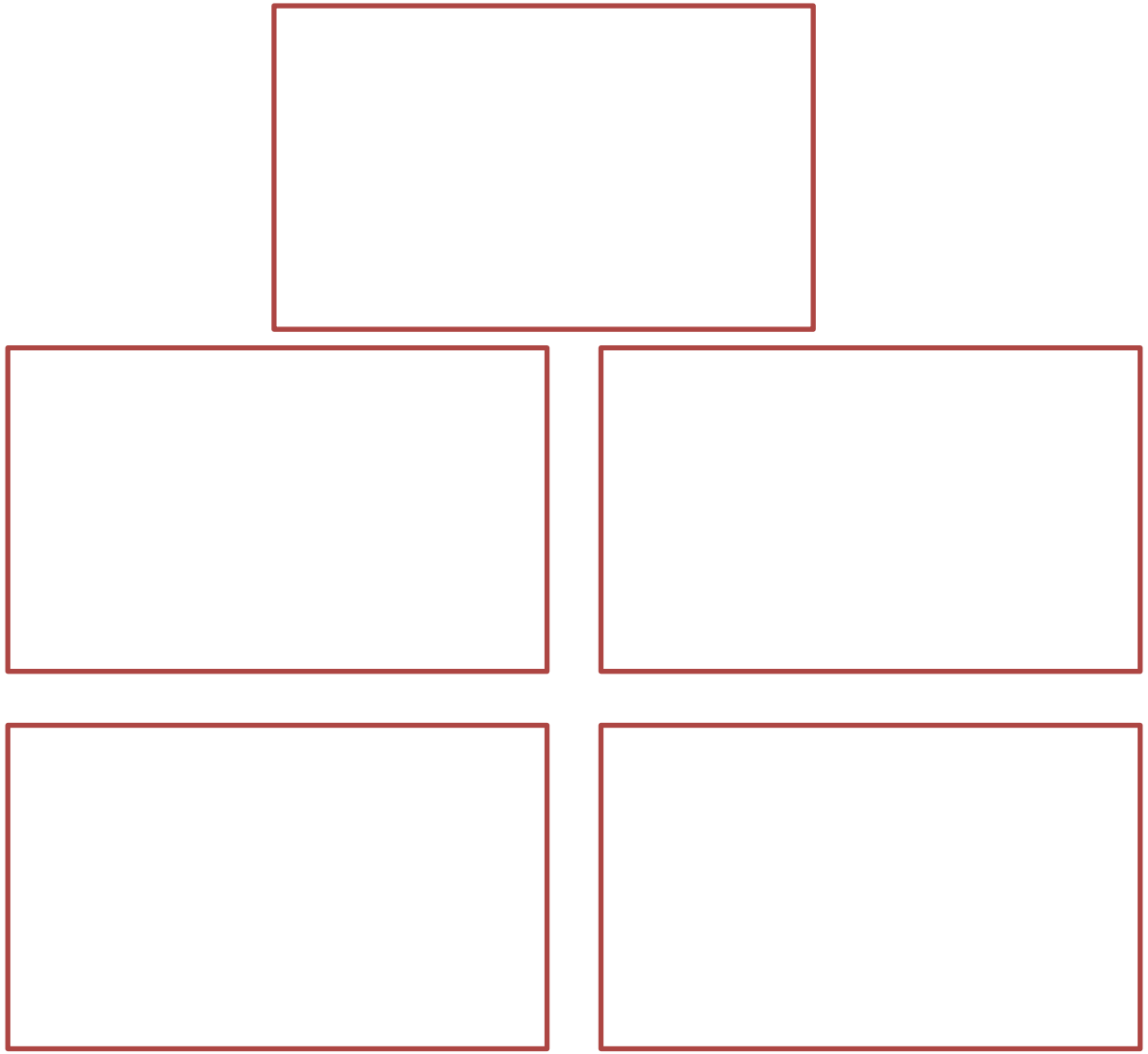
- MUC-*Message Understanding Conference* (final da década de 80)
  - Utilizava EI para automatizar tarefas de analistas de governo. Ex.: Digitalização de jornais com possíveis ligações ao terrorismo.

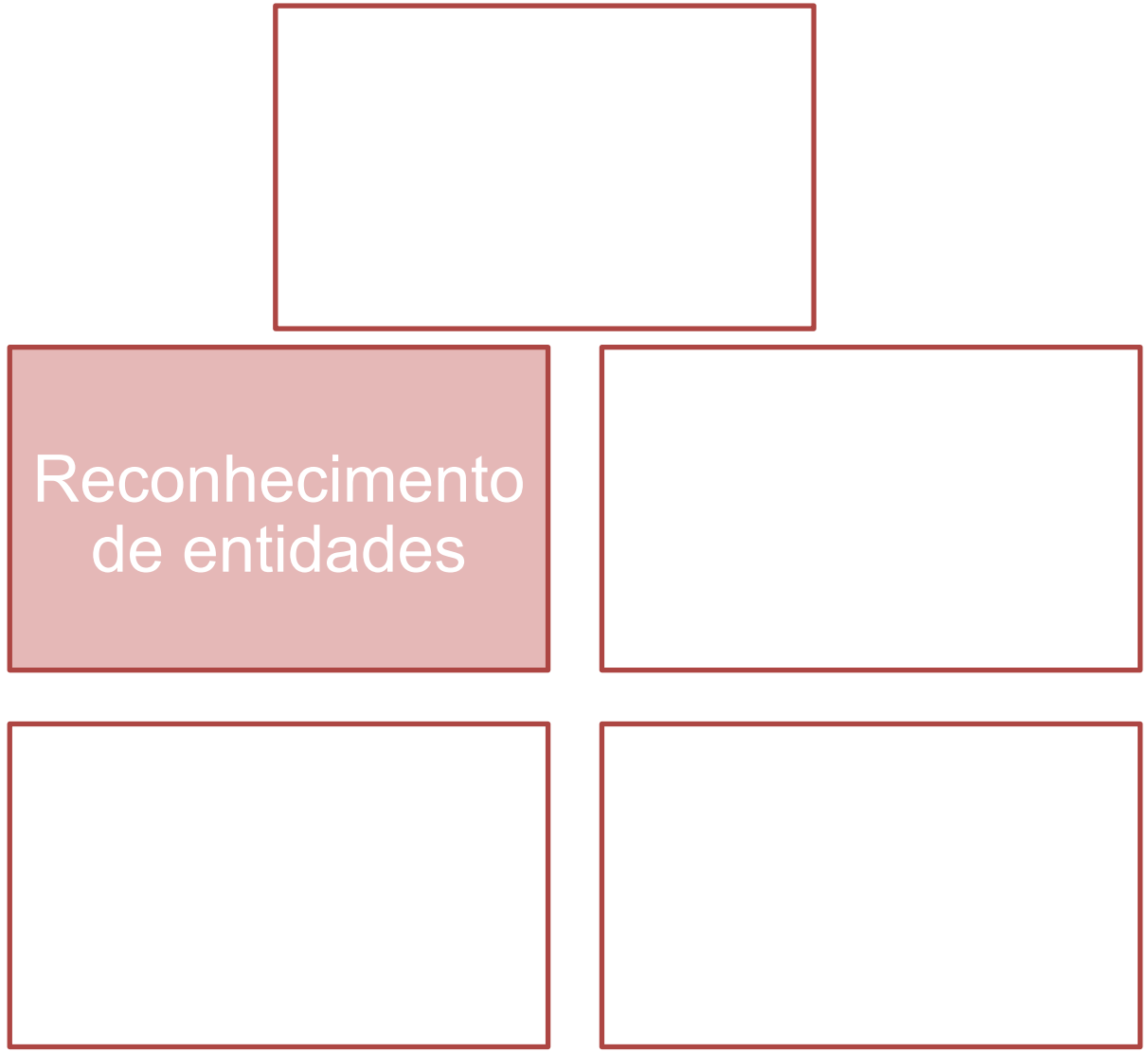
Por que EI é difícil?

- Linguagem Natural é difícil de tratar automaticamente
  - É muito flexível , ou seja, há várias formas para expressar uma única informação;
  - É ambígua, ou seja, a mesma sentença pode ter significados diferentes;
  - É dinâmica, ou seja, novas palavras podem ser introduzidas na língua, palavras existentes podem ganhar novos sentidos.

TIPOS DE TEXTO:

- **Estruturado**
  - Formato pré-definido e rígido
- **Não-Estruturado**
  - Livre
  - Sentenças em alguma linguagem natural
- **Semi-estruturado**
  - Frequentemente em estilo telegráfico
  - Formatação não segue regras rígidas




Reconhecimento de entidades

RECONHECIMENTO DE ENTIDADES:

- Detecção e classificação de entidades como:
  - Organizações (IBM Corporation)
  - Pessoas (Charles)
  - Nomes de lugares (New York)
  - Expressões temporais (6 de jun. 1911)
- Pode também extrair informação descritiva das entidades nos documentos.
  - Sexo
  - Nacionalidade
  - Posição

RECONHECIMENTO DE ENTIDADES:

- Detecção e classificação de entidades como:
  - 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.

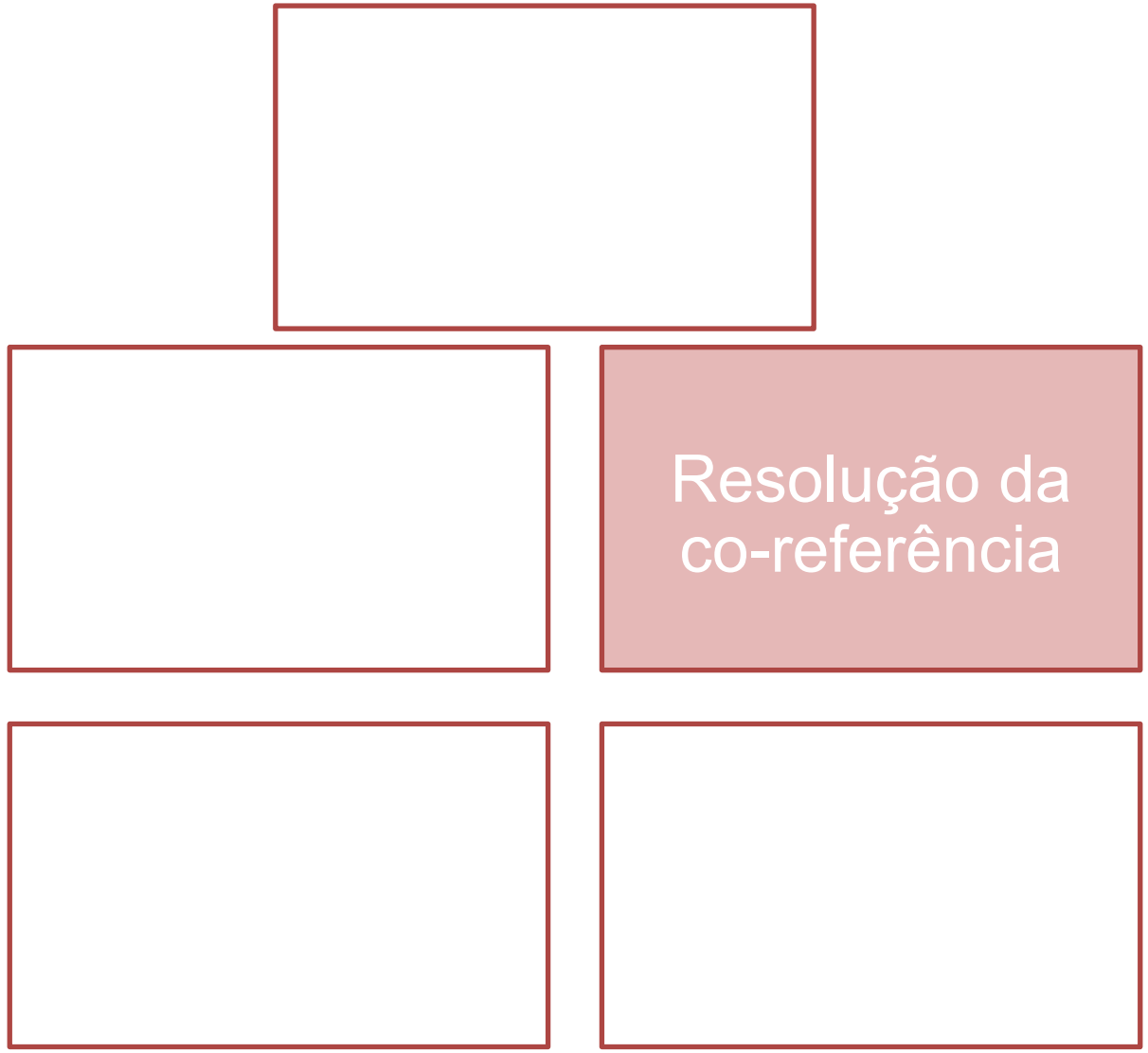
RECONHECIMENTO DE ENTIDADES:

- Detecção e classificação de entidades como:
  - 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.

Person
Date
Location
Organization

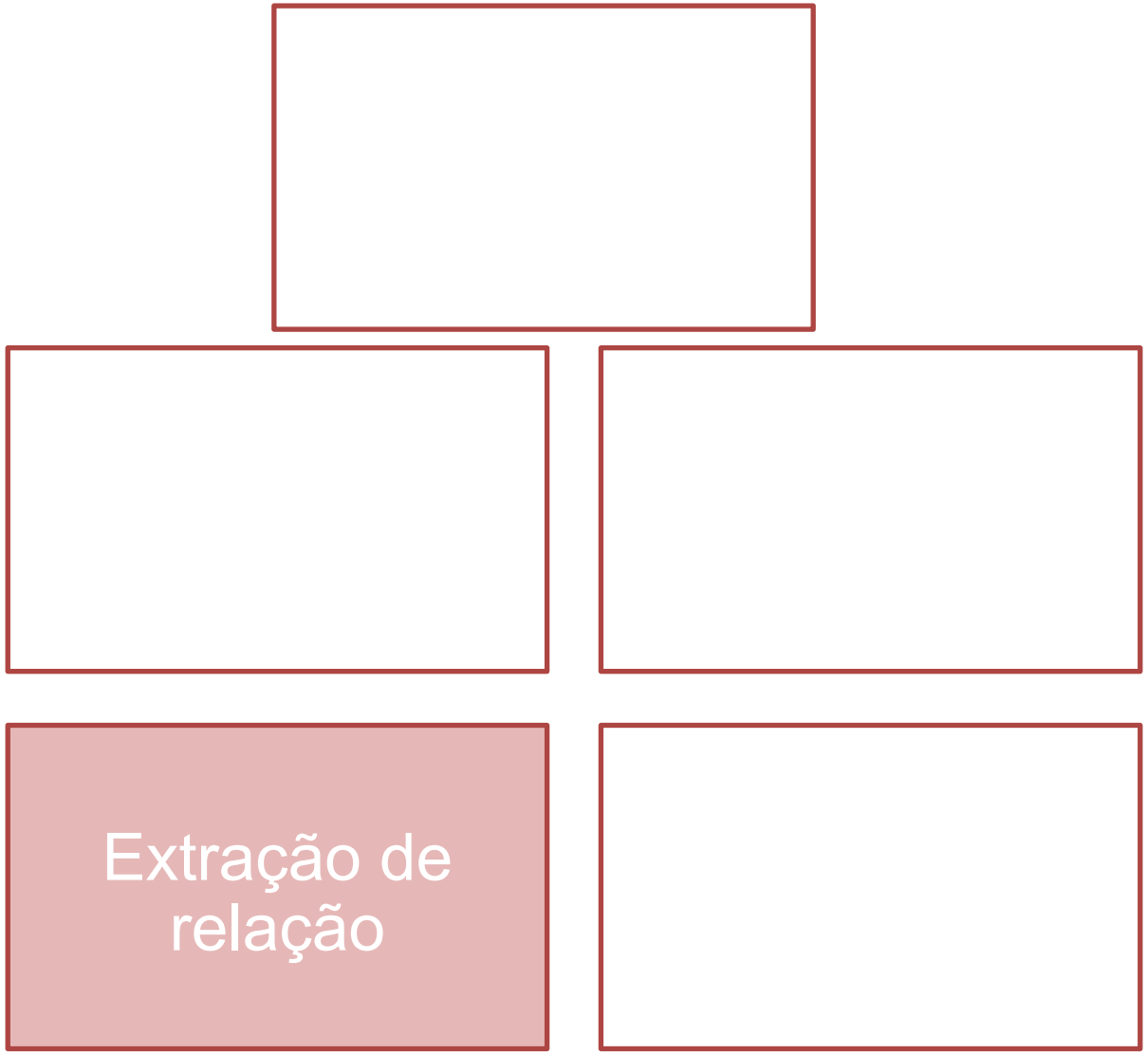
RECONHECIMENTO DE ENTIDADES:

- Usos
  - Entidades mencionadas podem ser indexadas
  - O sentimento pode ser atribuída a empresas ou produtos
  - Muitas relaçoes são asociações entre entidades.


Resolução da co-referência

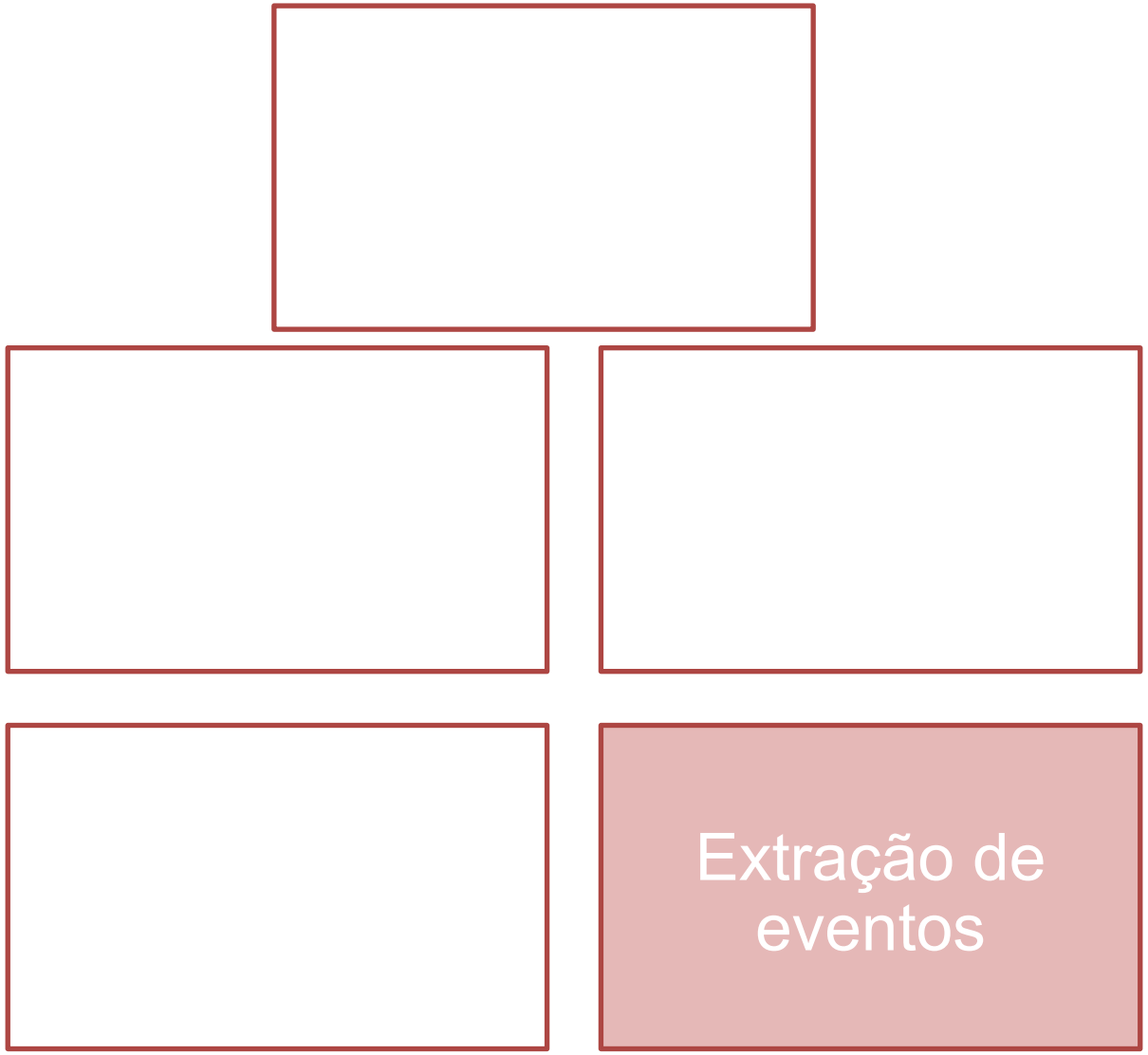
RESOLUÇÃO DA CO-REFERÊNCIA:

- Identificação de varias menções da mesma entidade no texto.
  - Caso que uma entidade é referida pelo nome: '*General Electric*' e '*GE*'.
  - Pronominal é quando a entidade é referida com um pronome: '*John bought food. But he forgot to buy drinks.*'
  - Nominal, no caso de uma entidade é referido com uma frase nominal: '*Microsoft revealed its earnings. The company also unveiled future plans.*'
  - Implícito: Berlusconi has visited the place of disaster. [He] flew over with a helicopter.


Extração de relação

EXTRAÇÃO DE RELAÇÃO:

- Detecta e classifica relações predefinidas entre entidades.
  - EmployeeOf(Steve Jobs,Apple)
  - LocatedIn(Smith,New York)
  - SubsidiaryOf(TVN,ITI Holding)


Extração de eventos

EXTRAÇÃO DE EVENTOS:

- Refere-se à tarefa de identificar eventos em texto livre e obter informação detalhada e estruturada sobre eles identificando: quem fez o quê a quem, quando, onde, através de que métodos e porque.
- '*Masked gunmen armed with assault rifles and grenades attacked a wedding party in mainly Kurdish southeast Turkey, killing at least 44 people.*'
  - Perpetradores (*masked gunmen*)
  - Numero de mortos e feridos (*at least 44*)
  - Armas (*rifles and grenades*)
  - Localização: (*southeast Turkey*)

TIPOS DE SISTEMAS PARA EI:

- **Baseados em PLN**
  - Extrair informações de textos em linguagem natural (livre)
  - Padrões linguísticos
- **Wrappers**
  - Principalmente para textos estruturados e semi-estruturados
  - Formatação do texto, marcadores, frequência estatística das palavras.

PROCESSAMENTO DE LINGUAGEM NATURAL:

- Utilizado no tratamento de documentos com pouco ou nenhum grau de estruturação.

- PLN caracteriza-se pela análise e manipulação ou codificação de informações expressas em língua natural a fim de encontrar os dados relevantes a serem extraídos.

PROCESSAMENTO DE LINGUAGEM NATURAL:

- **Processo de extração**
  - Extração de fatos (unidades de informação)

    Através da análise local do texto

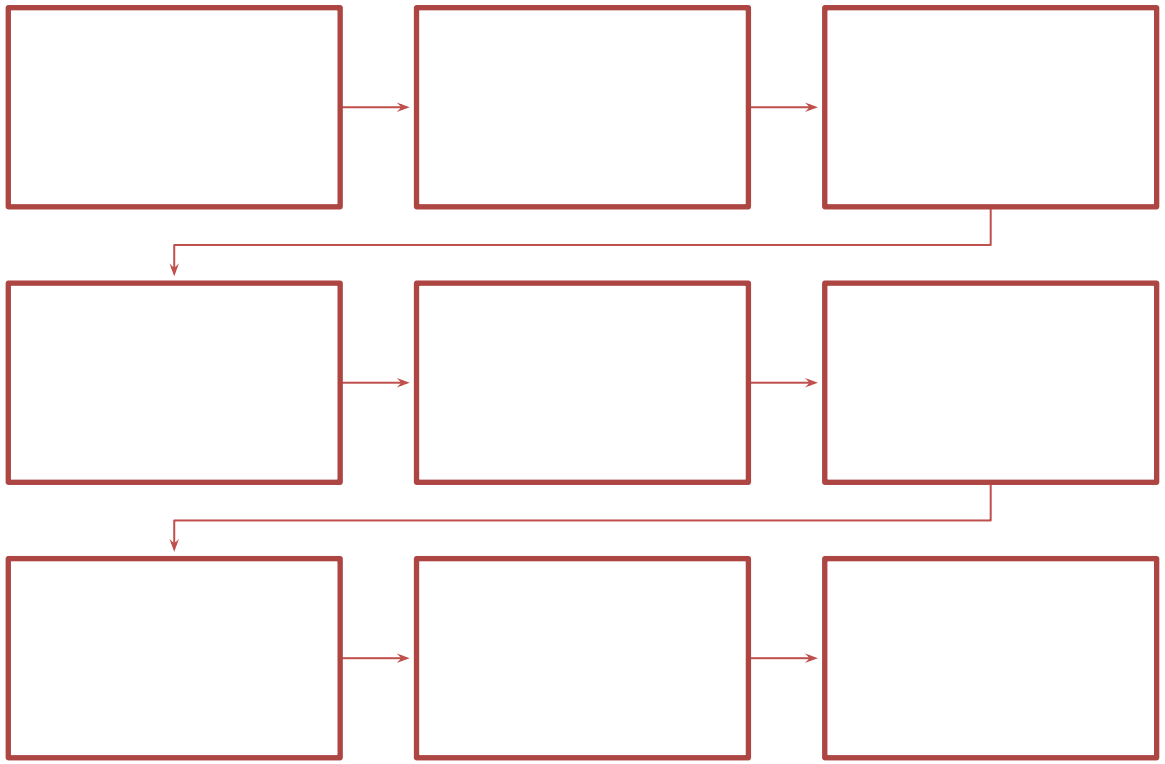
  - Integração e combinação de fatos

    Produção de fatos maiores ou novos fatos

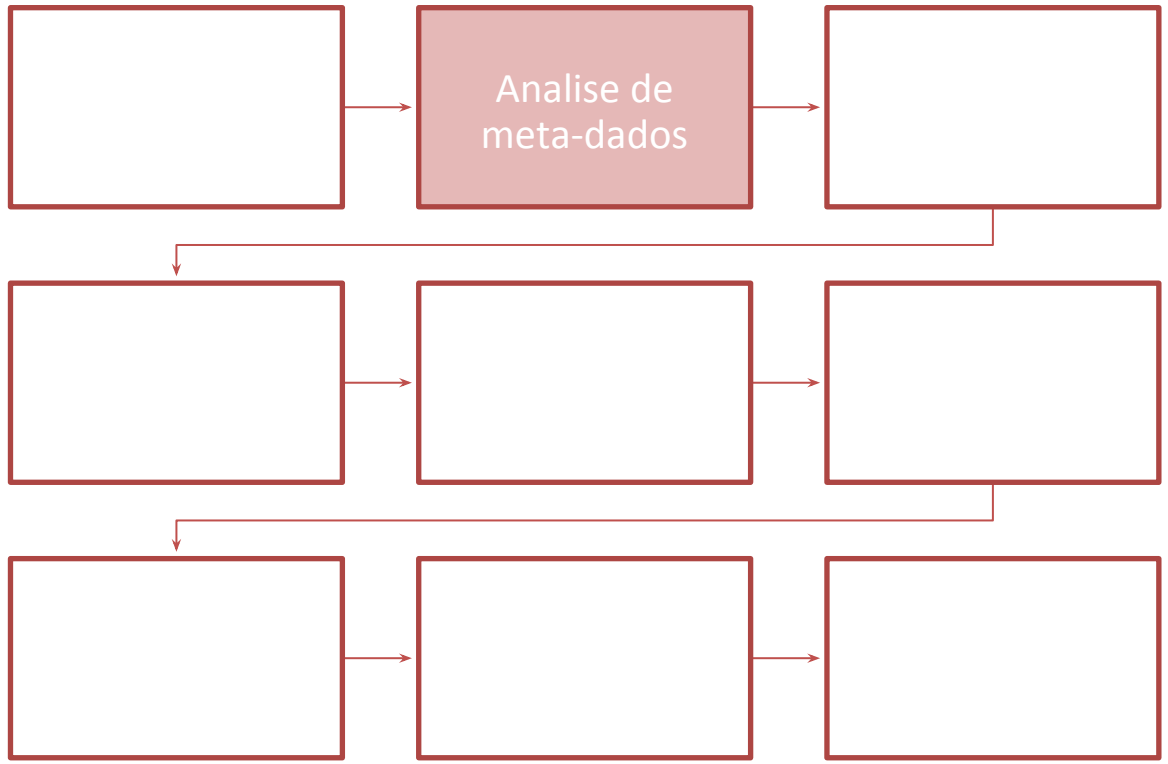
  - Estruturação de fatos relevantes

    Padrão de saída

Arquitetura de um sistemas para EI baseados em PLN

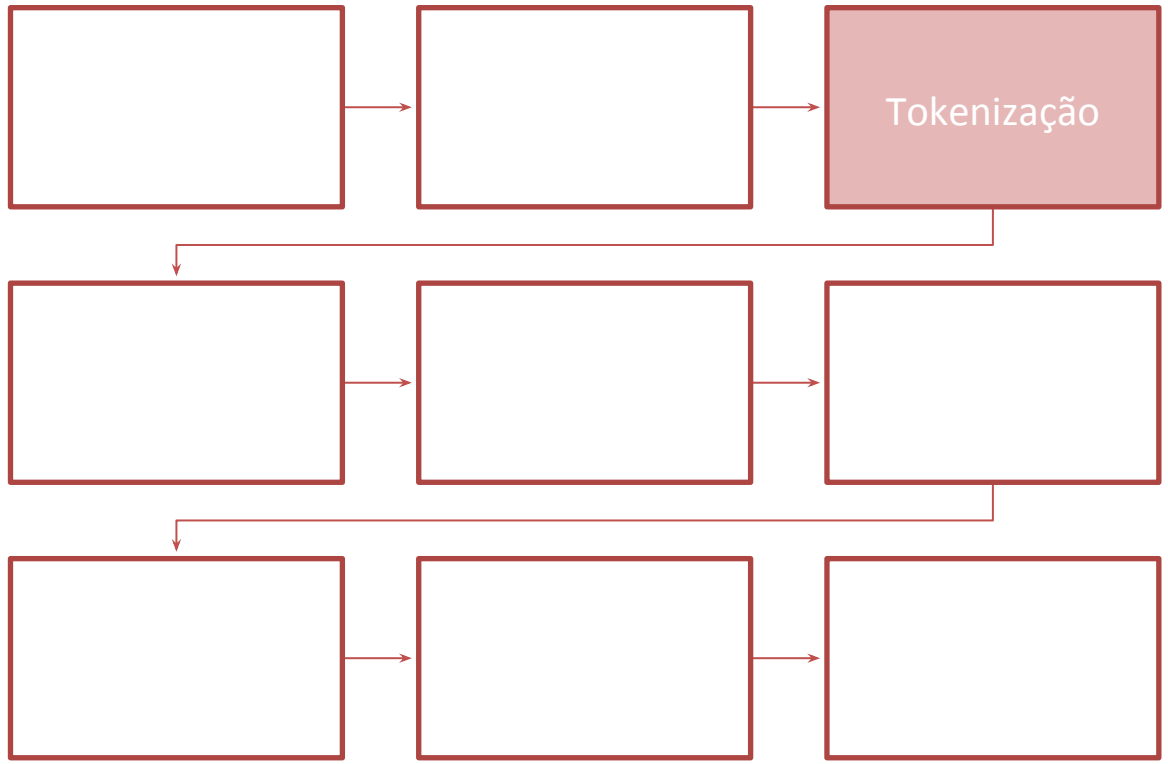


COMPONENTES DO SISTEMA DE EI:



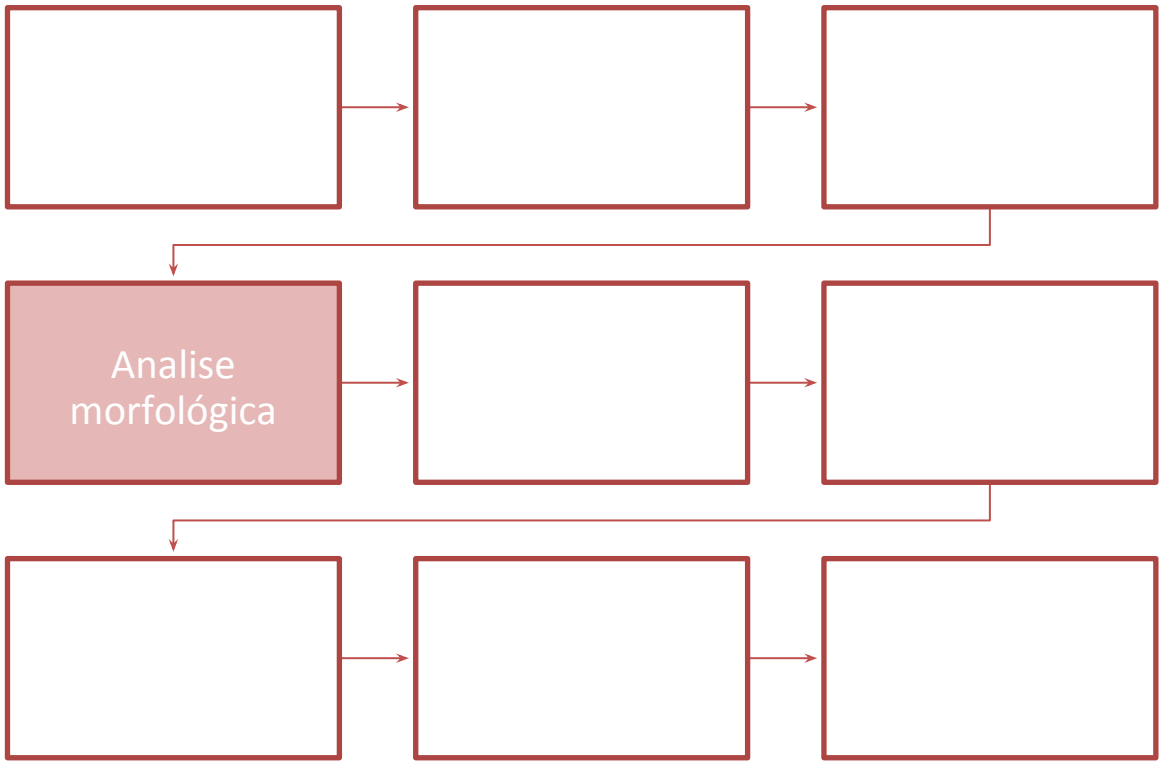
- Extração de títulos, corpo e data do documento.

COMPONENTES DO SISTEMA DE EI:



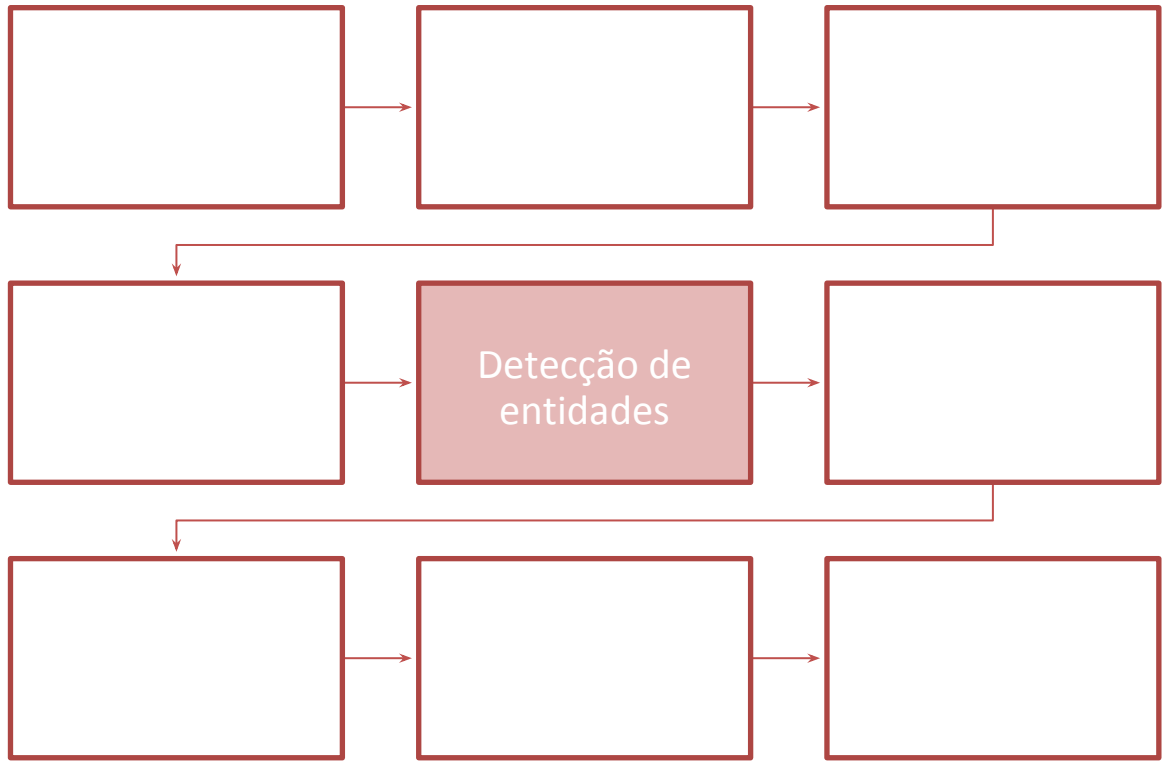
- Segmentação do texto em units o tokens e classificação de seu tipo.

COMPONENTES DO SISTEMA DE EI:



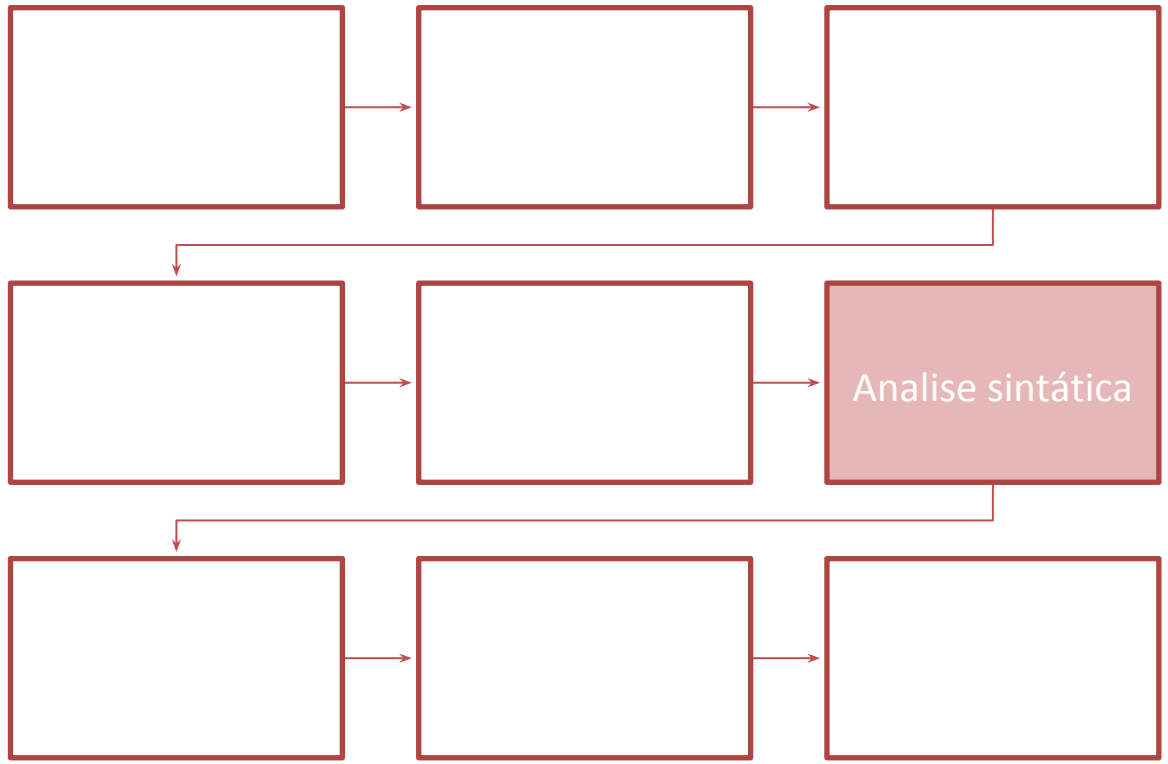
- Analisa a **classe gramatical** dos elementos:
  - Maria: substantivo próprio
  - comprou: verbo
  - um: artigo

COMPONENTES DO SISTEMA DE EI:



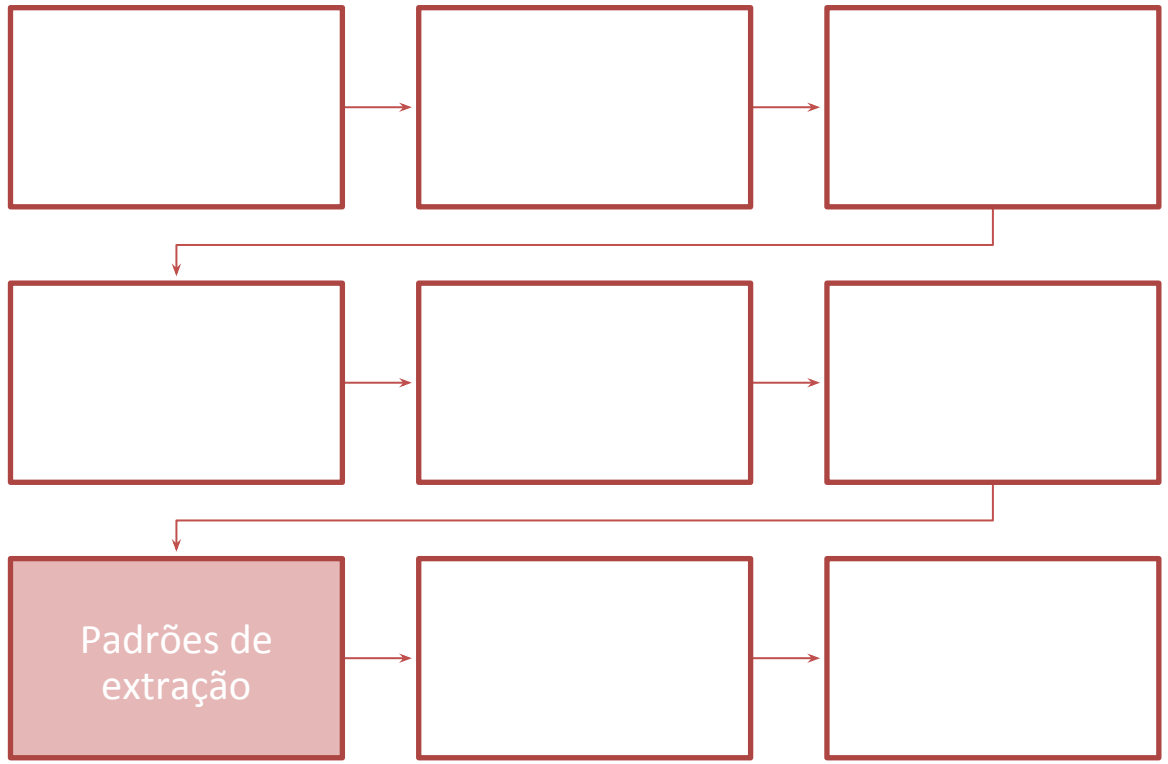
- Identifica nomes próprios.
- Itens que têm estrutura interna como da data e hora.
- Nomes são identificados por expressões regulares expressos em função das classes morfossintáticas (part-of-speech) e características sintáticas e ortográficas (letras maiúsculas) presentes nos termos.

COMPONENTES DO SISTEMA DE EI:



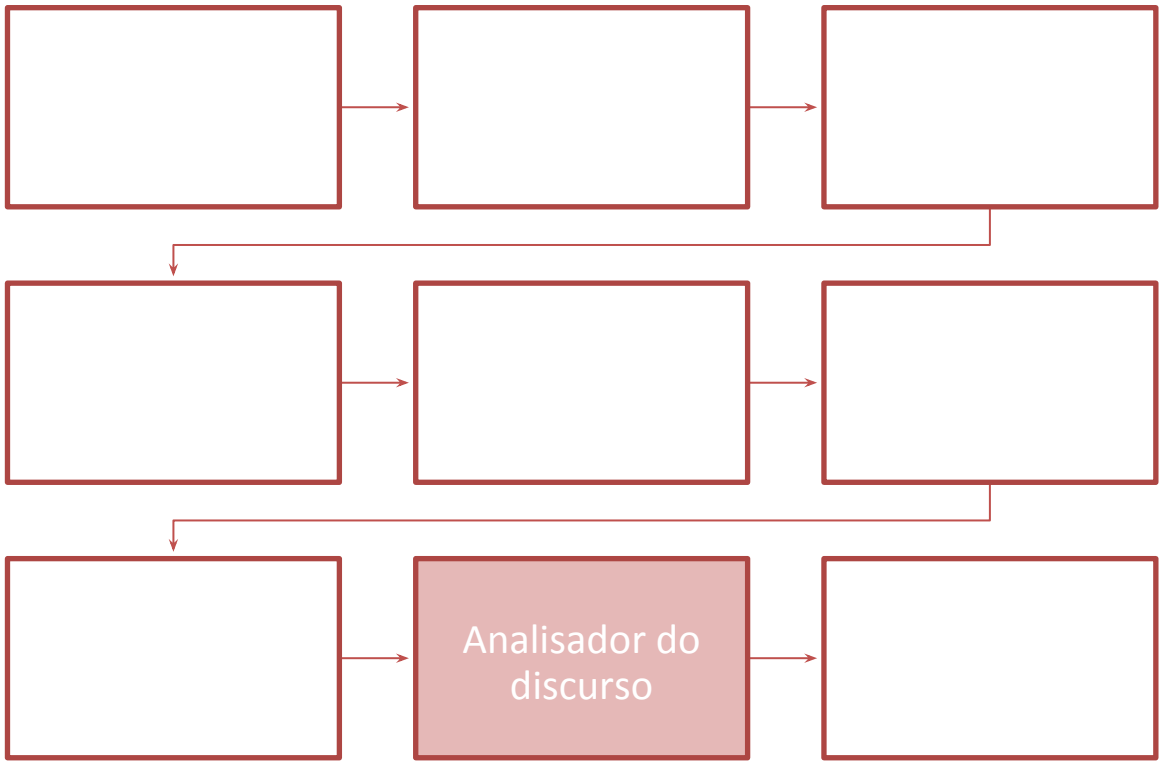
- A palavra não é estudada de forma isolada, pois ela mantém relação com outras palavras.
  - Júlia: Sujeito
  - Quebrou - Verbo transitivo direto
  - A cadeira: objeto direto
  - A: adjunto adnominal
  - Cadeira: núcleo do objeto direto

COMPONENTES DO SISTEMA DE EI:



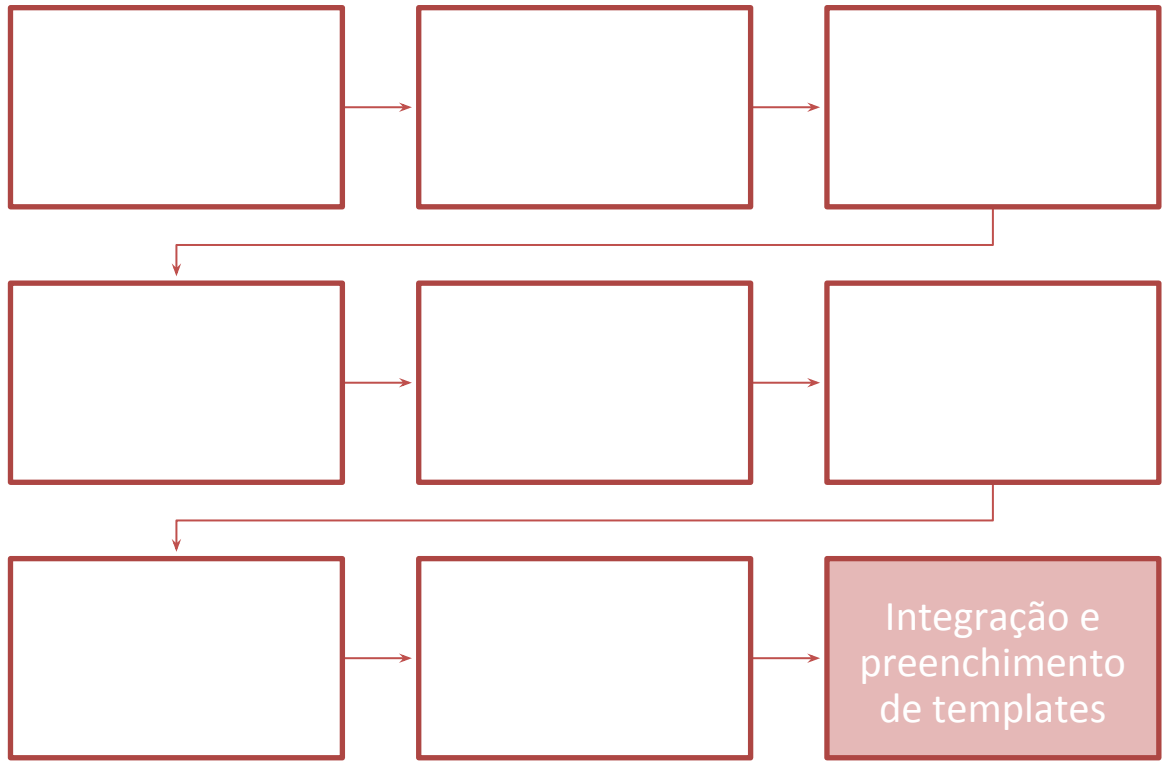
- Consiste na indução de um conjunto de regras de extração para o domínio tratado
- Esses padrões baseiam-se em restrições sintáticas e semânticas aplicadas as sentenças.

COMPONENTES DO SISTEMA DE EI:



- Análise de frases nominais, reconhece apostos e outros grupos nominais complexos.
- Resolução de correferência, identifica quando uma frase nominal se refere a outra já citada.
- Descoberta de relacionamento entre as partes do texto, para estruturar palavras do texto em uma rede associativa.

COMPONENTES DO SISTEMA DE EI:



- As informações são combinadas.
- Os templates são preenchidos com as informações relevantes ao domínio.

WRAPPERS:

- Maior desenvolvimento da WEB nos anos 90;
- Necessidade de sistemas mais eficientes com capacidade suficiente para extrair informação dos textos da WEB;
- Extraem a informação de documentos para preencher templates;
- Podem ser construídos de forma manual ou automática.

CONSTRUÇÃO MANUAL:

- Baseada em engenharia do conhecimento
  - Construção manual de regras de extração
  - Padrões de extração são descobertos por especialistas após examinarem o corpus de treinamento

- Vantagens
  - Boa performance dos Sistemas

- Desvantagens
  - Processo de desenvolvimento trabalhoso
  - Escalabilidade
  - Especialista pode não estar disponível

CONSTRUÇÃO AUTOMÁTICA:

- Aprendizagem de máquina
- Utiliza algoritmos de Inteligência Artificial
- Uma quantidade de documentos é utilizada no treinamento e geração das regras
  - Treinamento do sistema para novos textos

- Interação com o usuário pode ser feita
  - Aprende regras com a interação com o usuário

- Tempo menor de desenvolvimento

- Menor precisão nos resultados

Aprendizagem de Máquina

- Vantagens
  - Mais fácil marcar um corpus do que criar regras de extração
  - Menor esforço do especialista
  - Escalabilidade

- Desvantagens
  - Esforço de marcação do corpus de treinamento

Técnicas de Extração usadas em Wrappers

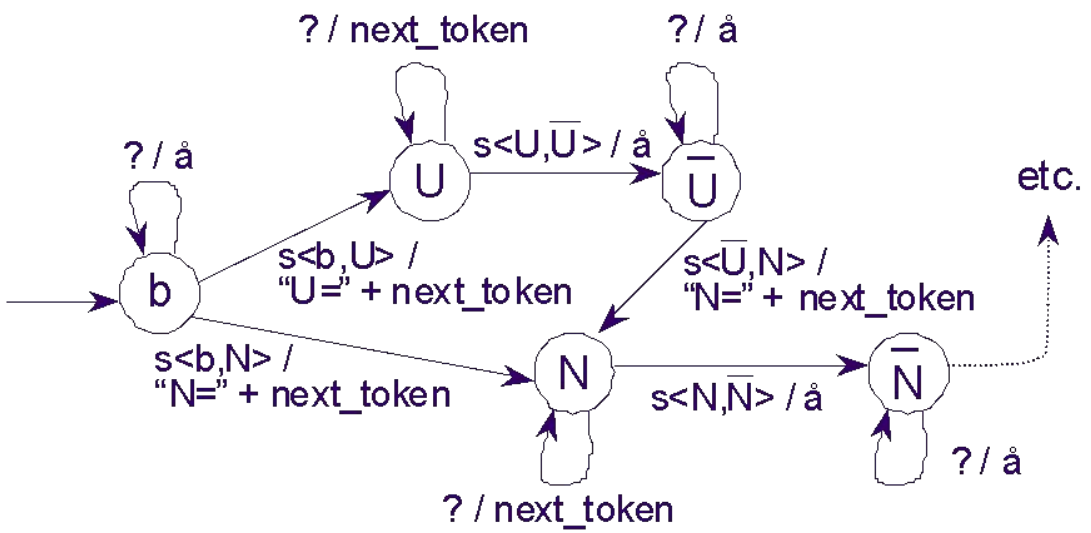
- Autômatos Finitos
- Casamento de Padrões
- Classificação de Textos
- Modelos de Markov Escondidos

Autômatos Finitos

- Regras de extração na forma de autômatos finitos
- Definidos por:
  - Estados que “aceitam” os símbolos do texto que preenchem algum campo do formulário de saída,
  - Os estados que apenas consomem os símbolos irrelevantes encontrados no texto.
  - Os símbolos que provocam as transições de estado
- Textos estruturados e semi-estruturados
  - Delimitadores, ordem dos elementos

Exemplo de Autômatos Finitos

```
<LI> <A HREF="…"> Mani Chandy </A>, <I>Professor of Computer
    Science</I> and <I>Executive Officer for Computer Science</I>
…
<LI> Fred Thompson, <I>Professor Emeritus of Applied Philosophy and
    Computer Science</I>
```



Key

- ? : wildcard
- U : state to extract URL
- Ū : state to skip over tokens until we reach N
- N : state to extract Name
- N̄ : state to skip over tokens until we reach A
- s<X,Y> : separator rule for the separator of states X and Y
- etc.

Casamento de Padrões

- Aprendem regras na forma de expressões regulares
- Expressões regulares que “casam” com o texto para extrair as informações
- Textos livres, estruturados e semi-estruturados
  - Delimitadores, padrões regulares (Ex. data, CEP)

Casamento de Padrões

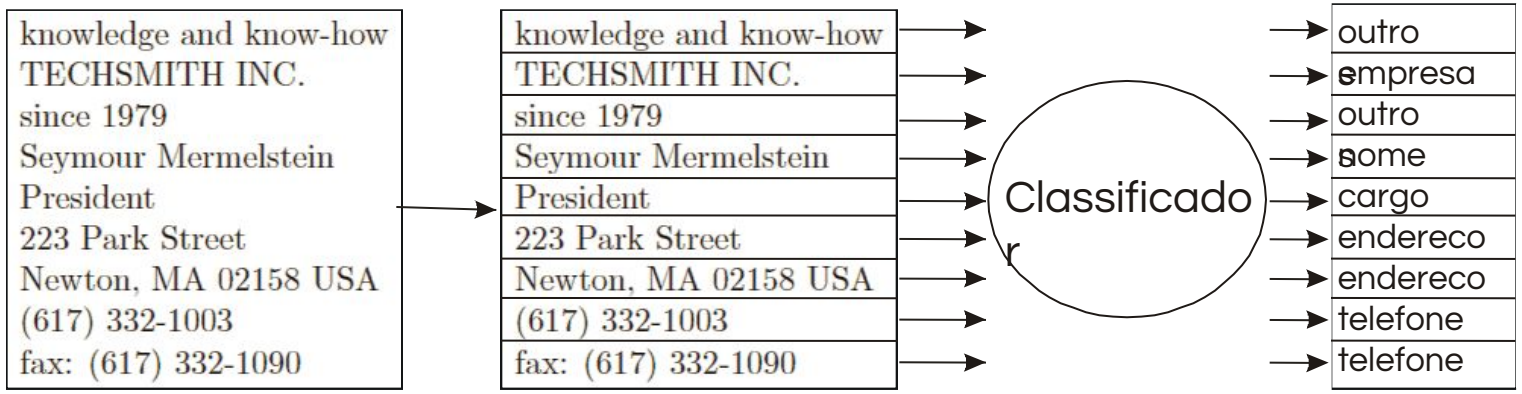
Padrão :: * (*Digit*) BR * $ (*Number*)
Formulário:: Aluguel {Quartos $1} {Preço $2}

Capitol Hill – 1 br twnhme. fplc D/W W/D.
Undrgrnd pkg incl $675 3 BR, upper flr
of turn of ctry HOME. incl gar, grt N. Hill
loc $995. (206) 999-9999
 (This ad last ran
on 08/03/97.) <hr>

Classificação de textos

- Dividem o texto de entrada em fragmentos candidatos a preencher algum campo do formulário de saída.
- Classificam os fragmentos com base em suas características
  - posição
  - número de palavras
  - presença de palavras específicas
  - letras capitalizadas
- Desvantagem
  - Classificação local independente para cada fragmento.

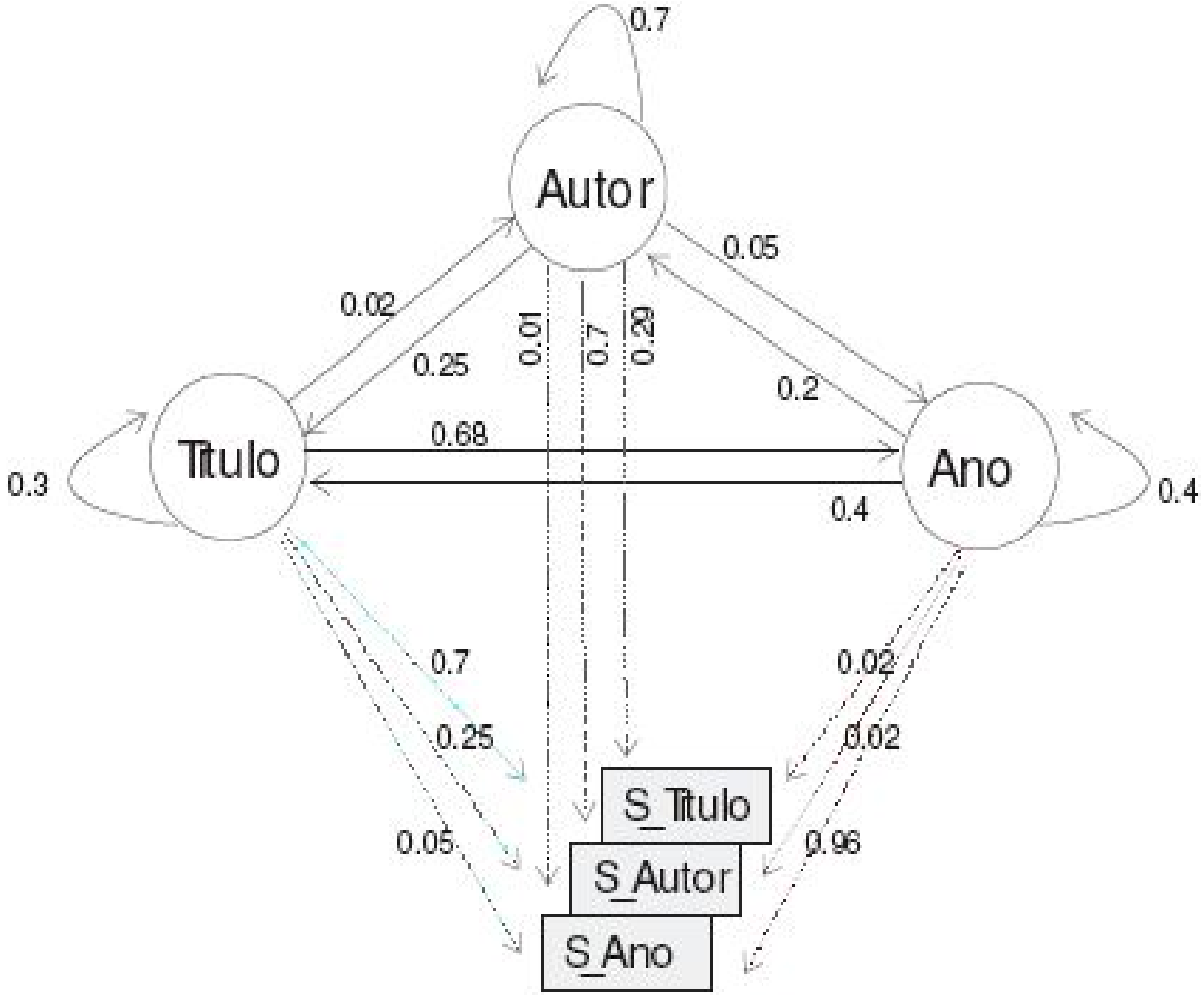
Classificação de textos



Modelos de Markov Escondidos (HMM)

- Um HMM é um autômato finito probabilístico que classifica seqüências de entrada
- Processo de classificação
  - Retorna a seqüência de campos com maior probabilidade para uma sequencia de fragmentos de entrada
- Vantagem
  - Realizar uma classificação ótima para a seqüência completa de entrada.

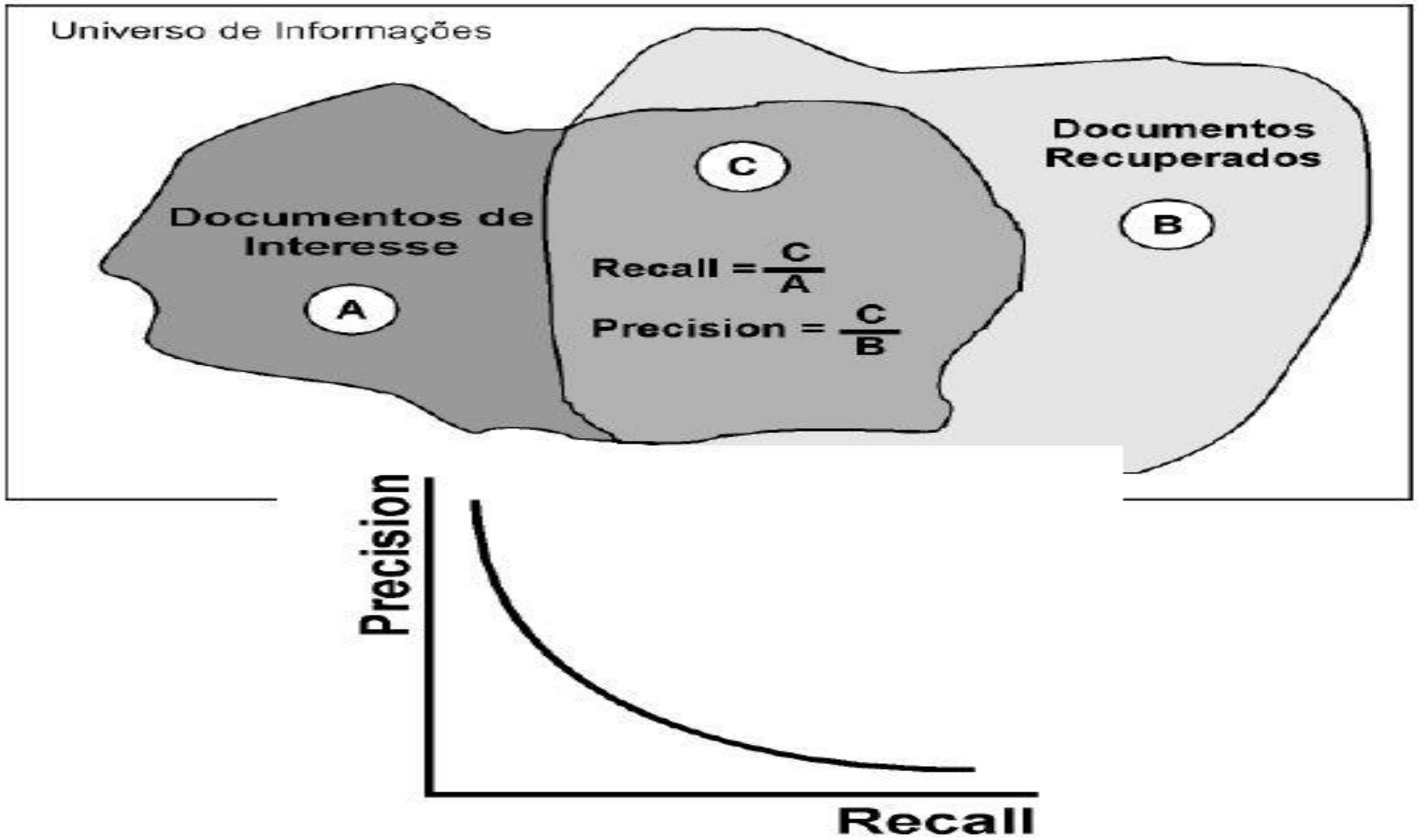
Exemplo


Autor
Titulo
Ano
S_Titulo
S_Autor
S_Ano
0.7
0.05
0.02
0.25
0.01
0.7
0.29
0.2
0.68
0.3
0.4
0.4
0.7
0.25
0.05
0.02
0.02
0.96

Avaliação em EI

- Informações extraídas X Informações desejadas
- $precisão = \frac{TP}{TP+TN}$
- $cobertura = \frac{TP}{TP+FN}$
- $F_{Measure} = \frac{(\beta^2+1)*cobertura*precisão}{(\beta^2)*Precisão+cobertura}$

Avaliação em EI



EI EM MÍDIAS SOCIAIS:

- Mídias sociais resultou em novas formas de expressão individual e comunicação.
- Enviar informações através da mídia social, incluindo blogs, fóruns Web, e micro-blogging serviços como Facebook ou Twitter.
- Permite escrever mensagens de texto curtas para comunicar, comentar e conversar sobre eventos atuais de qualquer espécie, eventos atuais, política, produtos, etc.
- Análise automatizada de conteúdo de mídia social, incluindo a extração de informações a partir da mídia social.

EI EM MÍDIAS SOCIAIS:

- Extração de informações de mídia social é mais desafiador do que clássico EI.
- Os principais problemas do processamento de conteúdo de mídia social são:
  - Os textos são tipicamente muito curto.
  - Textos são ruidosos e escrita em um ambiente informal.
  - Elevada incerteza da fiabilidade da informação transmitida na mensagens de texto.

Trabalhos de EI em Social Medias

- Locke e Martin apresenta resultados de ajuste de um classificador baseado em SVM para classificar pessoas, locais e nomes de organizações no Twitter.
- Benson apresenta uma abordagem baseada no CRF para extrair eventos de entretenimento.
- Liu propõe NER para segmentação e classificação de tweets combinando KNN e CRFs.

EXTRAÇÃO DE INFORMAÇÃO ABERTA:

- aborda relações ilimitadas, não requer exemplos de treinamento e abrange domínios genéricos, como a Web.
- Suas premissas são
- Independência de domínio;
- Extração não supervisionada;
- Escalabilidade para grandes quantidades de texto.
- O grupo de pesquisa KnowItAll da Universidade de Washington é o pioneiro do novo paradigma da EIA, que opera de uma maneira totalmente independente de domínio e na escala da Web.

EXTRAÇÃO DE INFORMAÇÃO ABERTA:

- A Extração de Informação tradicional opera em um pequeno conjunto de relações bem definidas e requer grandes quantidades de dados de treinamento para cada relação, ou seja, requer cada relação que se deseja extrair seja especificada como exemplo para as extrações. Esse paradigma é mais apropriado para extrações onde o número de relações é pequeno e o custo é baixo para rotular dados de treinamento

EXTRAÇÃO DE INFORMAÇÃO ABERTA:

- Quando o número de relações é massivo, e as relações não podem ser pré especificadas, a EIA é necessária, pois utiliza um extrator independente de domínio que escala para a Web e não precisa de nenhuma entrada humana. A força dos sistemas da EIA é o processamento eficiente, bem como a habilidade de extrair relações ilimitadas. Entretanto, comparada à EI tradicional, sua cobertura é baixa.

FERRAMENTAS PARA EI:

- **TextRunner**: primeiro sistema da EIA, desenvolvido por Banko e Etzioni. O sistema faz uma única passagem em um corpus de texto desestruturado e extrai uma grande quantidade de tuplas relacionais, sem requerer nenhuma entrada humana.
- **Wanderlust:** desenvolvido por Akbik e Bross, Resumidamente, rotula ligações entre as palavras de uma frase. As palavras ligadas possuem alguma dependência gramatical. Depois, essas ligações são analisadas e, a partir disso, as triplas são montadas.
- **WOE:** Wikipedia-based Open Extractor foi desenvolvido por Wu e Weld. Esse sistema utiliza correspondentes heurísticos entre valores de atributos de infoboxes daWikipédia e sentenças correspondentes para construir seus dados de treinamento, ou seja, cria paralelos entre relações extraídas do texto com relações da Wikipédia.

FERRAMENTAS PARA EI:

- **ReVerb:** é um programa desenvolvido por Fader e Etzioni. Esse programa usa um novo modelo para identificar e extrair automaticamente argumentos e relações expressas por verbos em sentenças em inglês. Ele se mostrou melhor que o TextRunner e o WOE. Além de corrigir problemas das versões anteriores e possuir melhor desempenho, o ReVerb mais que dobra a precisão e a cobertura.

- **R2A2:** é a segunda geração dos sistemas da EIA. Esse sistema é a combinação do ReVerb com o ArgLearner, um identificador de argumentos para extrair melhor argumentos para as relações baseadas em verbo.

FERRAMENTAS PARA EI:

- **OLLIE:** Open Language Learning for Information Extraction foi desenvolvido por Mausam et al., um sistema melhorado que alcançou alto rendimento ao extrair relações mediadas por substantivos, adjetivos e mais. Encontra mais extrações corretas em comparação ao ReVerb e aoWOEPARSE, mas também perde algumas extrações encontradas pelo ReVerb.
- **Kraken:** Desenvolvido por Akbik e Löser, o Kraken é um sistema projetado especialmente para capturar fatos n-ários, mas é vulnerável a ruídos.
- **ClausIE:** um extrator de relações e seus argumentos de textos em linguagem natural. O sistema foi desenvolvido por Del Corro e Gemulla. De acordo com os autores, o ClausIE é uma abordagem baseada em "cláusulas". Uma cláusula é uma parte de uma sentença que consiste de um sujeito (S), um verbo (V), e opcionalmente um objeto indireto (O), um objeto direto (O), um complemento (C), e um ou mais advérbios (A).

FERRAMENTAS PARA EI:

- **CORE:** utiliza uma série de técnicas do PLN para extrair informações de sentenças em Chinês.

- **DepOE:** é um sistema multilíngue baseado em análise (parsing) de dependência. O DepOE usa um analisador baseado em regras para realizar extrações em inglês, espanhol, português e galego.

- **GATE:** primeira versão lançada em 1996, trabalha com recursos (resources) que podem ser do tipo Visual Resources (VR), Language Resources(LR) e Processing Resources(PR).
  - Donwload
  - Tutorial

- **Stanford CoreNLP:** é uma estrutura integrada que fornece um conjunto de ferramentas de análise de linguagem natural.
  - Download

FERRAMENTAS PARA EI:

- Algumas bibliotecas em Python:

- pdfrw;
- Slate;
- PDFQuery;
- PDFMiner;

- PyPDF2;
- Para Java:

- iText

REFERÊNCIAS:

- Uma Abordagem de Aprendizagem Híbrida para Extração de Informação em Textos Semi-Estruturados. Eduardo F.A. Silva, Flávia A. Barros & Ricardo B. C. Prudêncio
- Schneider O. M., Rosa, L.J., Processamento de Linguagem Natural (PLN), http://moschneider.tripod.com/pln.pdf
- Aranha C., Passos E. A Tecnologia de Mineração de Textos, PUC-RIO
- Bulegon H., Moro M. C. C., Text Mining and Natural Language Processing in Discharge Summaries, PPGTS,PUCPR

REFERÊNCIAS:

- Pires. J.C. Batista, Extração e Mineração de Informação Independente de Domínios da Web na Língua Portuguesa, UFGO.

- Zambenedetti.C., Extração de Informação sobre Base de Dados Textuais, UFRGS.