Um Sistema Para Geração Automática de Questões

Luís Gustavo T. Cordeiro

Departamento de Informática e Estatística - Universidade Federal de Santa Catarina
Florianópolis, SC - Brasil

luis_gtc@hotmail.com

Abstract*. This paper briefly describes some basic techniques of natural language processing that can be used in systems for text analysis and question generation through sentences in a given language. This work is based on Portuguese language and, although some information will be useful for other languages, specific techniques for other languages are not presented. The most popular approaches are described and a basic model of question generation will be detailed. Some examples and results are presented for better understanding of the developed model.*

Resumo. Este artigo descreve brevemente algumas técnicas básicas de processamento de linguagem natural que podem ser utilizadas em sistemas para a análise de texto e geração de questões através de sentenças em determinada linguagem. O trabalho é baseado na língua portuguesa e, apesar de algumas informações também serem úteis para outros idiomas, não são apresentadas técnicas específicas de outras linguagens. As abordagens mais conhecidas são descritas e um modelo básico de geração de questões será detalhado. Alguns exemplos são apresentados para melhor entendimento, bem como resultados obtidos pelo modelo desenvolvido.

1. Introdução

A geração de questões é um ramo bem específico da área de processamento de linguagem natural, que por sua vez faz parte do imenso tema da inteligência artificial. Yao (2010) define o termo como uma tarefa que une os esforços das tarefas de entendimento de linguagem natural e geração de linguagem natural. De forma simples, é uma tarefa conhecida como um mapeamento *text-to-text*. Uma das formas de se realizar esse processo de geração de questões pode ser observada no modelo descrito em Cordeiro (2016). Para entendermos como funciona o sistema desenvolvido para geração automática de questões nos formatos Cloze e *WH-Question* no estilo Múltipla Escolha, primeiramente devemos entender os conceitos mais básicos sobre o assunto.

A forma como uma questão é apresentada ao leitor pode ser diferente dependendo do objetivo que se deseja obter e o contexto em que ela se encontra. Os formatos mais comuns em testes escolares são os de múltipla escolha, onde uma pergunta é realizada e são oferecidas diversas opções e o leitor decide qual delas se adequa melhor como a resposta. Smith e Avinesh (2010) apontam como sendo um formato interessante, pois possibilita a avaliação automatizada dos testes.

O formato de múltipla escolha pode ser combinado com outros tipos de questões, pois é apenas um modo de apresentar opções de resposta ao leitor. O sistema desenvolvido

em Cordeiro (2016) utiliza desse formato em conjunto com as *WH-Questions* e questões de tipo Cloze. Este último também é muito utilizado e bem conhecido como as questões de "preencher o vazio", onde uma sentença é apresentada com uma ou mais expressões removidas para serem completadas. Já as WH-Questions são perguntas que usamos naturalmente no dia a dia para solicitar uma informação específica. O termo é derivado do inglês e representa o conjunto de perguntas que iniciam com as expressões "quem", "quando", "onde", "o que", "de quem", "por que" e "qual". Além dessas, existem diversas outras formas de se apresentar uma questão, como as de ligação de conceitos; verdadeiro ou falso; sim ou não; e somatório; dentre outras das quais este artigo não apresentará em mais detalhes.

Um dos conceitos mais utilizados atualmente em trabalhos da área de processamento de linguagem natural é o corpus. Segundo a definição de Sinclair (2004), um corpus é uma coleção de textos de uma determinada linguagem em formato eletrônico, selecionados de acordo com um critério externo para representar uma linguagem como fonte de dados para pesquisa linguística. Em geral, quanto maior um corpus, maior a sua representatividade, seguindo a lógica probabilística de que uma quantidade maior de exemplares aumenta a chance de palavras e expressões mais raras serem encontradas no conjunto.

Diversos autores como Gasperin e Lima (2001) e Souza e Felippo (2010) apresentam suas ideias sobre características específicas que devem ser abrangidas pelos corpora para que possam ser considerados adequados para estudo, como padronização de um tamanho mínimo aceitável e outros conceitos de diversidade de temas. Estes temas não fazem parte do foco deste artigo e não serão abordados.

As técnicas utilizadas para processamento de texto estão diretamente relacionadas aos tipos de análise textual descritos por em Müller (2003). Segundo o autor, "um sistema de processamento de linguagem natural é abordado do ponto de vista da análise do conhecimento morfológico, sintático e pragmático". Sendo assim, surgiram técnicas como a tokenização, o *stemming* e a lematização, que tratam da identificação das palavras e suas formações morfológicas. As técnicas de etiquetagem morfossintática e da resolução de referências abrangem a análise sintática, identificando o tipo de cada palavra e sua relação com outras partes do texto. Ambas as técnicas são de grande utilidade para evitar erros básicos na geração de alternativas para as questões de múltipla escolha, como será visto mais à frente. Por último apresentamos o reconhecimento de entidades mencionadas: um processo de identificação de alto nível de abstração capaz de classificar os sujeitos e objetos das sentenças em um grupo como "pessoa", "local" e "tempo", utilizado como base para o desenvolvimento do sistema em Cordeiro (2016).

2. Abordagens Para a Geração de Questões

Le et al. (2014) cita três abordagens utilizadas para o processo de geração de questões: sintática, semântica e baseada em *template*. As duas primeiras abordagens seguem um fluxo de processamento muito semelhante, passando pela análise de uma sentença, removendo um conceito alvo, adicionando uma palavra chave da questão e convertendo o verbo para o formato gramaticalmente correto. A diferença entre a abordagem sintática e semântica está na etapa de transformações da sentença e nas regras que definem as modificações necessárias para tais, através da análise textual no nível correspondente.

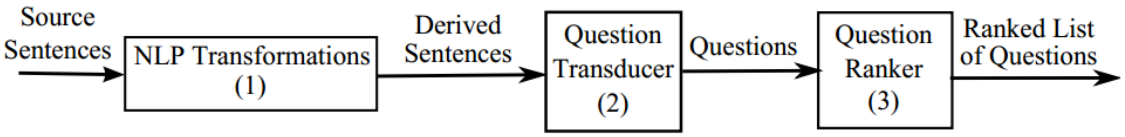
Tanto a abordagem semântica como a baseada em *template* existe uma dependência crescente de um domínio. É preciso ter um conhecimento prévio do tema abrangido pelo texto para definir adequadamente as regras de transformação, limitando seu uso. A utilização de ontologias na abordagem semântica possibilita a geração de questões através de análises mais profundas que se encontram diretamente ligadas ao texto escrito, podendo se utilizar de relações diferentes entre elementos, seus tipos e outras relações.

Os *templates* são modelos de sentenças desenvolvidos manualmente que servem como uma lista de verificações. Se uma sentença encaixar em um modelo, podemos gerar uma questão previamente definida para tal modelo, substituindo variáveis no texto. Tal método possibilita a geração de questões com parafraseamento, ou seja, as questões não refletem exatamente o texto da sentença geradora.

2.1. Processo de QG

O processo de geração de questões pode ser visto de forma genérica em poucos passos superficiais, porém, dependendo da abordagem utilizada, as etapas são ajustadas para realizar alguns detalhes específicos de cada abordagem. Heilman (2011) descreve um gerador de *wh-questions* com um fluxo simplificado de etapas (Figura 1) seguindo um modelo conhecido como *overgenerate-and-rank*, onde as sentenças são simplificadas, transformadas em questões e avaliadas com uma pontuação a fim de remover as inadequadas.

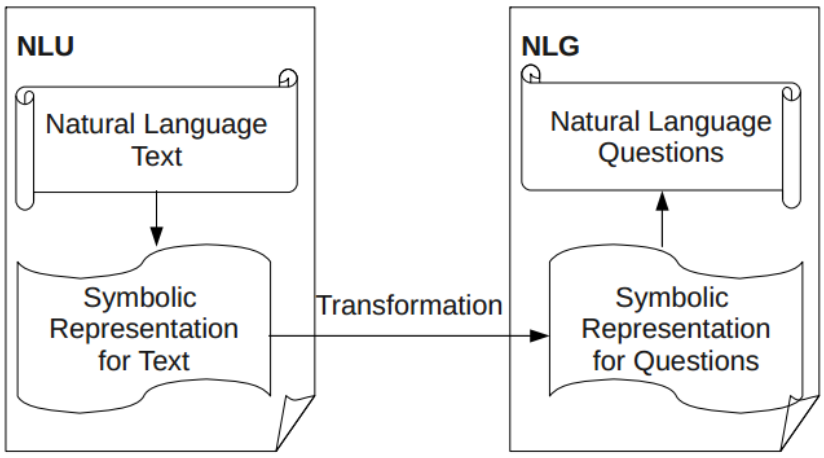
Figura 1 - *Framework* Para Geração de *WH-Questions*



Fonte: Heilman (2011, pg 45)

Em Yao, Bouma e Zahng (2012) foi definido um sistema baseado na abordagem semântica contendo apenas três etapas: transformação de texto em uma representação simbólica, conversão em uma representação simbólica para a questão e a transformação final da questão para linguagem natural, como pode ser visto na Figura 2.

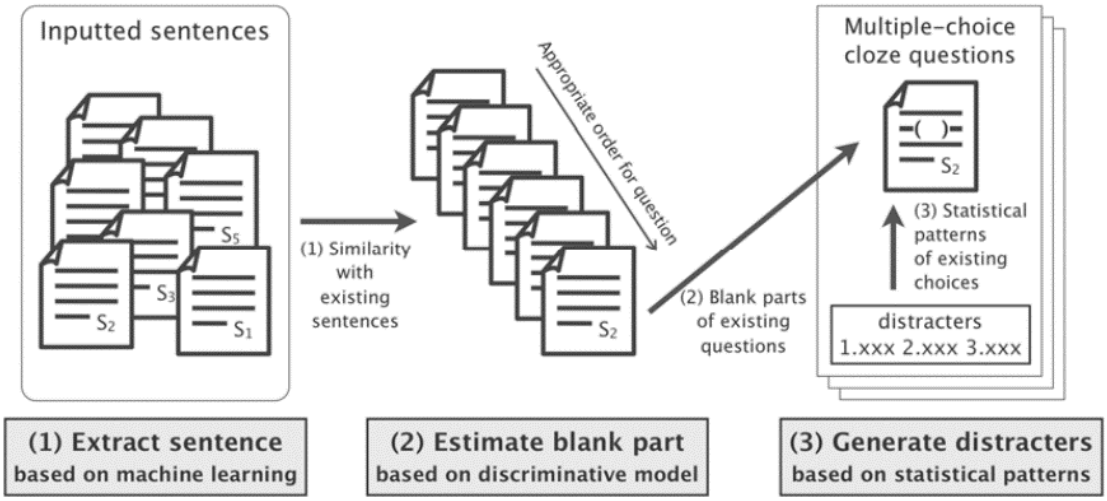
Figura 2 - Relação Entre NLU e NLG no Processo de Geração



Fonte: Yao, Bouma e Zahng (2012, pg 12)

Um processo simplificado para a geração de questões do tipo Cloze múltipla escolha também foi utilizado em Goto et al. (2009) com uma abordagem estatística para seleção dos componentes da questão. O processo também segue três etapas: seleção de uma sentença, estimativa de uma parte da sentença para remoção e geração dos distrativos (alternativas de resposta da questão), demonstrado na Figura 3.

Figura 3 - Processo de Geração de Questões Cloze Múltipla Escolha



Fonte: Goto et al. (2009, pg 2)

3. Sistema de Geração de Questões Cloze e WH em PT-BR

A seguir serão apresentados o modelo de geração de questões utilizado em Cordeiro (2016) e o modo de funcionamento do sistema desenvolvido.

3.1. Modelo

O sistema funciona através de um aplicativo Java de linha de comando com a leitura de um arquivo formato texto (.txt) do qual são extraídas as sentenças e gerando um arquivo texto de saída com uma lista das questões geradas. Existe a opção de escolher entre os tipos Cloze e *wh-question* ou ambos, podendo também ativar ou desativar a geração de uma quantidade de distrativos desejada.

Seguindo a classificação definida por Graesser, Rus e Cai (2008), o sistema tem como propósito o monitoramento do entendimento do tópico apresentado no texto por parte do leitor, com a intenção de possibilitar uma futura adequação para uma plataforma de estudos. O tipo de informação tratado é bastante abrangente e adequa-se em várias classes, abordando definições, especificações e complemento dos conceitos apresentados pelo texto, porém evita questões que requerem comparações e opiniões.

A análise textual é realizada através de um modelo probabilístico gerado com a utilização de algoritmos de aprendizado de máquina na ferramenta Apache OpenNLP com um conjunto de corpora da língua portuguesa do projeto Floresta Sintá(c)tica. A utilização dessa abordagem evita a criação manual de diversas regras para analisar as sentenças e suas estruturas sintáticas. Enquanto para a geração das questões é utilizada uma abordagem mista entre a sintática e semântica, baseando-se principalmente na identificação e classificação de entidades mencionadas no texto. O sistema segue um pouco a ideia de overgenerate-and-rank, porém não foi desenvolvida a parte que realiza a poda dos resultados indesejados.

3.2. Funcionamento

O aplicativo recebe as entradas necessárias na linha de comando e inicia com o *parsing* dos parâmetros. Os principais valores recebidos são o caminho do documento de texto de entrada e o caminho do arquivo de saída, porém também é possível modificar os modelos que serão utilizados para a análise textual, os formatos de questão (Cloze e WH) e a quantidade de distrativos que serão gerados para cada questão.

Com os modelos na memória, o sistema segue com a leitura e análise do arquivo de entrada. O texto é dividido em sentenças, e cada uma delas passa pelo processo de tokenização, dividindo-a em diversos tokens, que em geral são palavras, e cada um é analisado tendo suas informações como a classe morfológica, grau e número armazenados em um objeto. Uma lista desses tokens é guardada para cada sentença, também em uma outra estrutura.

A partir desse ponto, o sistema possui as sentenças em um formato que possam ser passadas para análise com o objetivo de obter uma árvore sintática com a representação de suas partes (sujeito, objeto, adjunto, etc.) e uma lista com as entidades mencionadas (pessoa, tempo, local, organização, evento).

3.2.1. Geração de Questões Cloze

No caso da geração das questões Cloze, apenas as entidades mencionadas são utilizadas. O algoritmo de geração é bem simples, apenas removendo os tokens que fazem parte da entidade da sentença original e substituindo-os por uma parte vazia. Como exemplo vamos utilizar a sentença exemplo número 6 em Cordeiro (2016, pg. 85).

(S.6) Stefano Evodio Assemani trabalhou na Biblioteca Apostólica Vaticana como intérprete de línguas orientais.

Nessa sentença, o sistema identifica duas entidades mencionadas: “Stefano Evodio Assemani” e “Biblioteca Apostólica Vaticana” a partir das quais são geradas as questões A17 e A18, respectivamente, com a omissão dos termos.

(A.17) _____ trabalhou na Biblioteca Apostólica Vaticana como intérprete de línguas orientais.

(A.18) Stefano Evodio Assemani trabalhou na _____ como intérprete de línguas orientais.

3.2.2. Geração de *WH-Questions*

Esse tipo de questão é muito mais complexo de ser gerado, pois requer determinadas regras definidas com intuito de executar a transformação da sentença original em um formato interrogativo. Essas regras são complicadas de serem definidas e implementadas programaticamente, pois dependem do idioma, e do resultado da análise sintática para identificar corretamente as partes que serão removidas, reposicionadas ou adicionadas na sentença.

Atualmente o sistema leva em consideração apenas sujeitos, adjuntos adverbiais e objetos preposicionais para a geração das *wh-questions*. O algoritmo funciona procurando por partes na árvore sintática com uma das três classificações e então verifica se existe alguma entidade mencionada no trecho. Se existir, a wh-word é escolhida dependendo do tipo da entidade, conforme mostra o Quadro 1, caso contrário, a expressão é ignorada.

Quadro 1 - CONDIÇÕES PARA A DEFINIÇÃO DA WH-WORD

TIPO DA ENTIDADE	WH-WORD
Place (local)	Onde?
Person (pessoa)	Quem?
Time (tempo)	Quando?
Numeric (numeral)	Quantos?
Outros	O que?

Fonte: produzido pelo autor

As entidades de tipo numérico deveriam conter um tratamento especial, pois o ideal seria identificar a medida utilizada (tamanho, tempo, etc.) para modificar a *wh-word* e fazer perguntas do tipo "quanto tempo?" ou "quanto pesa?", porém não foi desenvolvido nenhum tratamento do tipo.

Para modificadores das *wh-words* foi implementada a adição de preposições quando a entidade faz parte de um adjunto adverbial ou um objeto preposicional. O algoritmo procura por preposições "por", "para", "com", "de" e suas variações ("pelo", "da", etc.) gerando questões do tipo "por quem?", "com quem?" e "para onde?", dentre outras. Como exemplo temos a sentença exemplo número 4 em Cordeiro (2016, pg. 84).

(S.4) The Magic Cloak of Oz é um filme dirigido por J. Farrell MacDonald.

Nesse exemplo, a entidade "J. Farrell MacDonald" é removida e define-se a wh-word "quem". Porém, a parte "por J. Farrell MacDonald" é classificada como sendo um adjunto adverbial preposicional, sendo assim o sistema identifica a preposição "por" e a adiciona como modificador da wh-word, resultando na questão V4.

(V.4) Por Quem The Magic Cloak of Oz é dirigido?

Um dos pontos levantados com alguns exemplos foi a necessidade de remover o complemento do sujeito mantendo apenas o verbo, pois quando a expressão era mantida a questão gerada não fazia sentido. Esse caso ocorre no exemplo anterior, mas será explicado em um outro caso, com a sentença exemplo número 2 (CORDEIRO, 2016, pg. 83).

(S.2) Arumana no Kiseki é um jogo de videogame lançado pela Konami em 1987.

No exemplo S2, o sistema identifica a entidade Arumana no Kiseki, porém a simples remoção do termo resultaria na questão "O que é um jogo de videogame lançado pela Konami em 1987?". Talvez a sentença ideal seria a substituição da wh-word "o que" por "qual" e a troca do artigo indefinido "um" pelo artigo definido "o", porém apesar de parecer simples nesse exemplo, não é uma tarefa trivial e foi optado pela remoção do complemento, mantando apenas o verbo, resultando na questão A6.

(A.6) O que é lançado pela Konami em 1987?

O último detalhe da geração de uma wh-question é o reposicionamento de adjuntos adverbiais no início da sentença para o final. Esse procedimento foi uma decisão tomada porque normalmente nesses casos a expressão é separada por vírgula e quando transformada a sentença para forma interrogativa, a questão não ficava muito natural. Esse caso pode ser visualizado na sentença exemplo número 7 em Cordeiro (2016, pg. 86) e demonstrado a seguir.

(S.7) Em 1898, Lucy Maud Montgomery voltou para Cavendish para viver com a avó viúva.

Nessa sentença foi identificada a entidade "Lucy Maud Montgomery", definindo a *wh-word* "quem" por sua classificação "pessoa". Porém, se seguida a regra básica de substituição obtemos a questão "Em 1898, quem voltou para Cavendish para viver com a avó viúva?", decidida como não sendo tão adequada quanto a questão A22, o sistema faz um reposicionamento do adjunto adverbial para o final da sentença, removendo a vírgula.

(A.22) Quem voltou para Cavendish para viver com a avó viúva em 1898?

3.2.3. Geração de Distrativos

Como as questões Cloze são baseadas nas entidades mencionadas, a geração de distrativos não é um processo muito complexo. Durante a geração das questões são armazenadas as entidades encontradas, sem repetições, sendo assim possível utilizá-los após a geração para escolher as alternativas para cada questão. Desta forma, o sistema é capaz de gerar os distrativos sem a necessidade da utilização de uma ferramenta externa, pois o próprio texto provê as alternativas. Porém, caso o texto de entrada seja muito pequeno e não possua muitas entidades de um determinado tipo, as alternativas geradas serão muito semelhantes na maioria das questões.

As questões geradas são armazenadas em uma estrutura de dados contendo também a sua resposta, ou seja, a entidade mencionada removida da sentença original, sendo assim possível obter suas informações. Isso é um fato importante, pois podemos selecionar para as alternativas apenas entidades de mesmo tipo, levando a um resultado mais satisfatório, ou seja, com menor chance de erros semânticos. Outro ponto levado em consideração para a seleção são as informações dos tokens da entidade mencionada, evitando erros gramaticais graves de concordância verbal, de gênero e de número.

Para as *wh-questions*, porém, o cenário é um pouco diferente. Como a geração é orientada por partes sintáticas, a remoção do texto em boa parte dos casos vai além da entidade mencionada. Por esse motivo, as alternativas geradas não encaixam adequadamente como sendo uma resposta viável em alguns casos e durante o desenvolvimento optou-se por manter essa forma simplificada nestes casos.

4. Conclusão e Trabalhos Futuros

Este artigo apresentou de forma breve e resumida o trabalho desenvolvido em Cordeiro (2016), desde a definição teórica básica, os conceitos de processamento de linguagem natural, passando por uma explicação mais detalhada do modelo desenvolvido e o funcionamento do sistema implementado para geração de questões. Neste último capítulo serão resumidos os resultados obtidos e os trabalhos futuros citados pelo autor.

O aplicativo obteve sucesso e está funcionando com bons resultados para a geração de questões tipo Cloze abertas, porém quando adicionada a opção de geração de distrativos, algumas alternativas não são adequadas, apesar de exceções, os resultados ainda podem ser considerados satisfatórios, pois cumpriram o objetivo de gerar questões válidas de forma automatizada. A geração de *wh-questions* é muito básica e simplificada, funcionando corretamente apenas em sentenças pequenas e diretas, sem uma estrutura sintática complexa.

Para melhoria dos resultados, o autor sugere a utilização de modelos probabilísticos mais precisos, diminuindo a repercussão de tais falhas para os algoritmos

de transformação de sentenças em questões. Além disso, também seria de grande utilidade a implementação de um processo para simplificação das sentenças, evitando diversos erros na geração de questões factoide. Também ajudaria a gerar resultados mais confiáveis a implementação do ranking dos resultados, possibilitando remover as questões inadequadas e menos naturais da lista final apresentada ao usuário.

As tarefas citadas anteriormente ajudariam na qualidade dos resultados finais, porém também seria interessante a utilização de outros artifícios para a geração de uma variedade maior de questões. Dois pontos levantados pelo autor nesse quesito são a resolução de referências e a utilização de paráfrases, possibilitando a geração de questões com uma apresentação diferente (ordem das palavras, sinônimos, etc.) com o mesmo significado, mas com um formato mais natural para o ser-humano. A falta de resolução das referências no sistema faz com que diversas possíveis questões sejam omitidas, pois pronomes de anáfora não são considerados entidades mencionadas. Outra opção para aumentar a variedade seria a implementação de novos algoritmos para geração de outros tipos de questões, como verdadeiro ou falso e somatórios.

O sistema desenvolvido apresenta bons resultados iniciais para aplicações voltadas a educação, porém ainda é preciso resolver muitos desses problemas para que os resultados se tornem confiáveis e de alta qualidade.

Referências

CORDEIRO, Luís G. T. **Geração Automática de Questões Através de Análise de Texto**. Florianópolis, SC, Brasil. 2016.

GOTO, Takuya; KOJIRI, Tomoko; WATANABE, Toyohide; IWATA, Tomoharu; YAMADA, Takeshi. **An Automatic Generation of Multiple-choice Cloze Questions Based on Statistical Learning**. 2009.

GRAESSER, Arthur C.; RUS, Vasile; CAI, Zhiqiang. **Question classification schemes**. In: Proc. of the Workshop on Question Generation. 2008.

HEILMAN, Michael. **Automatic Factual Question Generation from Text**. School of Computer Science. Carnegie Mellon University. Pittsburgh, PA, Estados Unidos. 2011.

MÜLLER, DANIEL N. **Processamento de Linguagem Natural**. Porto Alegre, Rio Grande do Sul, Brasil. 2003.

SINCLAIR, John. **Corpus and Text: Basic Principles**. Tuscan Word Centre. Developing Linguistic Corpora: a Guide to Good Practice. 2004.

SMITH, Simon; AVINESH P. V. S. **Gap-fill Tests for Language Learners: Corpus-Driven Item Generation**. In: Proceedings of ICON-2010: 8th International Conference on Natural Language Processing. Índia. 2010.

YAO, Xuchen. **Question Generation With Minimal Recursion Semantics**. 2010.

YAO, Xunchen; BOUMA, Gosse; ZHANG, Yi. **Semantics-based Question Generation and Implementation**. 2012.