# Mineração de Mapas Conceituais a partir de Textos em Português

**Camila Z. Aguiar[1], Davidson Cury[1]**

[1]Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
Vitória – ES – Brasil

{camila.zacche.aguiar, dedecury}@gmail.com

***Abstract.*** *The automatic construction of concept maps is a hot topic, especially when the map must represent, through concepts and relations, a complex and grammatically regulated text. Thus, we start from a detailed study on these approaches and from it, we propose in this article a new approach. To this end, we define a technological architecture that includes: (i) grammar rules and depth search for extraction of elements from the text, (ii) relevance computation of concept based on frequency and map topology, and (iii) graph-based method to summarize the map. The approach developed and preliminary results are presented.*

***Resumo.*** *A construção automática de mapas conceituais ainda é um assunto em voga, principalmente quando o mapa deve representar por meio de conceitos e relações um texto complexo e gramaticalmente regrado. Assim, partimos de um estudo detalhado sobre essas abordagens e, deste estudo, propomos neste artigo uma nova abordagem. Para esse fim, definimos uma arquitetura tecnológica que compreende: (i) regras gramaticais e busca em profundidade para a extração dos elementos a partir do texto, (ii) cálculo da relevância do conceito baseado em sua frequência e na topologia do mapa, e (iii) método baseado em grafos para sumarizar o mapa. A abordagem desenvolvida e resultados preliminares são apresentados.*

## 1. Introdução

Segundo Novak & Cañas (2008), conceitos e relações formam a base para o aprendizado e por isso, mapas conceituais têm sido fortemente utilizados em diversas situações e para diferentes finalidades, principalmente na educação. Mapas conceituais têm sido utilizados como recurso de aprendizagem, meio de avaliação, organização instrucional, representação cognitiva, elicitação ou compartilhamento do conhecimento. Ademais, podem representar a informação de documentos extensos, uma vez que uma representação gráfica mais dinâmica e flexível, em forma de conceitos e relações, é considerada mais fácil de ser construída, assimilada e compreendida do que um texto extenso e gramaticalmente regrado.

O procedimento padrão para a construção de um mapa conceitual envolve definir um tópico ou questão focal, identificar e listar os mais importantes conceitos relacionados ao tópico, ordenar os conceitos por ordem de relevância de cima para baixo no mapa e adicionar e rotular as frases de ligações entre os conceitos. Mesmo com a popularização de certas técnicas para a construção de mapas conceituais, sua construção ainda requer dedicação de tempo e esforço empenhado na identificação e estruturação do

conhecimento, especialmente quando a construção do mapa é realizada a partir do "zero", ou seja, quando seus elementos constituintes não são pré-definidos e precisam ser totalmente descobertos.

Com o objetivo de facilitar o processo de construção, várias abordagens tecnológicas têm sido propostas para auxiliar ou automatizar, de alguma forma, este processo. Neste artigo apresentamos nossa contribuição à educação, uma nova abordagem tecnológica para a construção automática de mapas conceituais a partir de textos científicos, em língua portuguesa. A abordagem compreende: (i) regras gramaticais e busca em profundidade para a extração dos elementos a partir do texto, (ii) cálculo da relevância do conceito baseado em sua frequência e topologia do mapa, e (iii) método baseado em grafos para sumarizar o mapa. A análise dos dados é apresentada, alcançando 0.68/0.38 e 0.41/0.19 de precision/recall para conceitos e relações.

Este artigo está estruturado como segue: na Seção 2 discutimos e apresentamos trabalhos relacionados à construção automática de mapas conceituais; a Seção 3 apresenta a proposta de uma nova abordagem; na Seção 4 mostramos os experimentos e resultados obtidos até o momento; e na Seção 5 são feitas algumas considerações preliminares.

## 2. Analisando o Estado da Arte

A partir de uma revisão da literatura sobre abordagens tecnológicas para construção automática de mapas conceituais entre os anos de 1994 e 2016, nas bases IEEE Xplore, ACM e Elsevier Science Direct (detalhada em [Aguiar et al. 2016]), aplicamos a categorização proposta em [Aguiar et al. 2016] para obter uma análise objetiva sobre essas abordagens. Aplicamos os seguintes filtros da categorização à revisão da literatura: *Style (Scientific), Structure (Unstructured) e Labeling (Present)* onde quatro trabalhos relacionados foram encontrados nesse contexto. Uma vez que as abordagens possuem o mesmo objetivo, automática construção de mapas a partir de textos, o que as difere é o processo realizado para a extração dos elementos do texto e a representação gráfica do mapa gerado.

A abordagem [Wang et al. 2008] gera mapas conceituais a partir de resumos em inglês. A abordagem utiliza análise morfológica e sintática, identificando os elementos com base na estrutura das frases e regras sintáticas. Aplica a normalização para corrigir erros ortográficos, depende da detecção de sinônimos e da resolução de anáfora. Usa análise estatística para verificar a relevância das proposições, fazendo uso da interação com o usuário para definir proposições incertas.

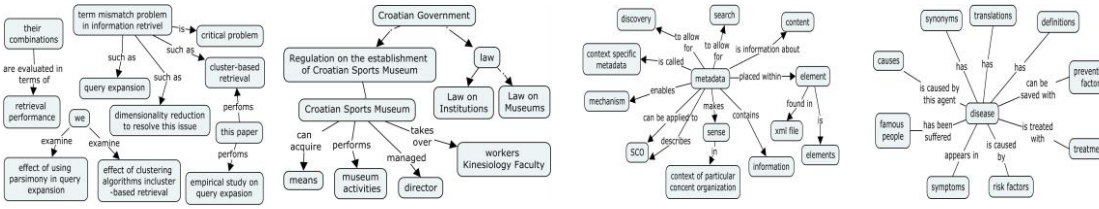
A abordagem [Zubrinić et al. 2015] gera mapas a partir de documentos legais em língua croata como um resumo do texto. Esta abordagem cria mapas hierárquicos de uma área específica usando um thesaurus do domínio. A partir de um corpus de domínio, os documentos são pré-processados e os metadados são mapeados. Usa técnicas linguísticas para lematização, reconhecimento de entidades, resolução de co-referência, análise léxica e sintática. Os conceitos são identificados a partir dos metadados e da frequência dos termos no texto. As proposições são extraídas a partir do padrão sujeito-predicado-objeto que contenha os conceitos identificados e relações estabelecidas entre os conceitos em um thesaurus. Uma estrutura de árvore formada por 25-30 conceitos das proposições é construída hierarquicamente atribuindo o título do texto ao nó raiz.

A abordagem [Zouaq & Nkambou 2009] gera mapas conceituais de textos em inglês como etapa intermediária para gerar uma ontologia. Para isso, utiliza técnicas

linguísticas de segmentação, normalização, análise estatística e sintática. Aplica o aprendizado de máquina para identificar palavras-chave e cria um mapa semântico de frases contendo essas palavras-chave. As triplas são extraídas a partir das regras sintáticas e dependências gramaticais entre as palavras na frase. Os padrões léxico-semântico interpretam essa estrutura para extrair conceitos e relações. Finalmente, realiza análise estatística para definir a relevância de conceitos e relações.

A abordagem [Villa et al. 2012] gera mapas conceituais de texto clínico em língua inglesa. Esta abordagem usa conceitos e uma ontologia para obter ricas informações sobre o domínio. O sistema pré-processa um conjunto de termos médicos compilados em uma lista e busca por termos do domínio no texto. O usuário escolhe um conceito e consultas são realizadas na base de conhecimento para recuperar informações sobre o conceito.

Olhando para os mapas gerados pelas abordagens (Figura 1), podemos observar: mapa fragmentado em porções [Wang et al. 2008]; rótulo de conceitos longo [Wang et al. 2008] e formado por pronome [Wang et al. 2008]; rótulo de relação ausente [Zubrinić et al. 2015] ou formado por preposição [Wang et al. 2008]; utiliza outras fontes de dados além do texto como ontologia [Zouaq & Nkambou 2009], base de conhecimento [Villa et al. 2012] e thesaurus [Zubrinić et al. 2015]; mapa criado a partir de um conjunto de documentos [Zouaq & Nkambou 2009], ou pequeno texto contendo algumas sentenças [Villa et al. 2012]; e mapa que representa apenas um domínio específico [Villa et al. 2012; Zubrinić et al. 2015].




[Wang et al 2008] [Zubrinic et al 2015] [Zouaq & Nkambou 2009] [Villa et al 2012]

**Figura 1. Mapa conceitual construído pelos trabalhos relacionados**

Analisando as abordagens apresentadas, observamos que nenhuma delas está direcionada à construção de mapas a partir de textos em português e pouquíssimas não são limitadas a um domínio específico do texto.

## 3. Uma Abordagem para Mineração de Mapas Conceituais

Nesta seção propomos uma abordagem para a mineração de mapas conceituais a partir de textos em idioma português, como uma extensão da abordagem [Aguiar et al. 2015] proposta em idioma inglês. A abordagem foi desenvolvida como um serviço web utilizando Java e *framework* Spring. O modelo conceitual é formado por onze etapas, iniciando na etapa de Preparação, com o recebimento da fonte de dados e finalizando na etapa de Sumarização, com a construção de proposições, na forma de conceito-relação-conceito. A Figura 2 apresenta a síntese desse processo.

A etapa de **Preparação** é responsável por receber e tratar a fonte de dados, ou seja, o texto em idioma português. Para isso, são usadas atividades para (i) eliminar marcadores de rótulos, referências, *tags* e estilo da fonte; e (ii) remover caracteres especiais.

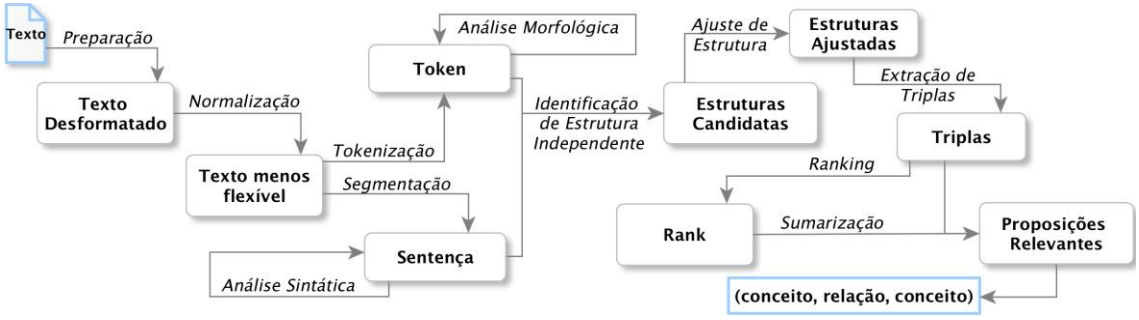



**Figura 2. Modelo Conceitual da Abordagem**

A etapa de **Normalização** altera a fonte de dados de modo a permitir a extração de informações mais inteligíveis. Para isso, são usadas atividades para (i) remover frases não-proposicionais; e (ii) resolver anáfora com a biblioteca OpenNLP[1]. *Stopwords* não são removidas nesta etapa, dado que a sua remoção influencia negativamente os processos posteriores.

As etapas de **Tokenização** e **Análise Morfológica** são realizadas em paralelo, com foco em termos individuais. Usamos os *tokens* que contêm *tags* de nome (NN), adjetivo (JJ), verbo (VB), determinante (DT), advérbio (RB) e preposição (IN). Para isso utilizamos a biblioteca OpenNLP[1] com o módulo Tokenizer treinado com o corpus Bosque[1] e módulo PosTagger, treinado com o corpus MacMorpho[2]. Como o corpus definem diferentes *tagset* e não podem ser combinados, uma conversão foi realizada.

As etapas de **Segmentação** do texto e **Análise Sintática** são realizadas em paralelo, com foco nas sentenças do texto. Para isso são utilizados sintagmas *clause level* contendo simples declarativa (S) e sintagmas *phrase level*, contendo frases nominais (NP), verbais (VP) e preposicionais (PP). Para isso utilizamos a biblioteca OpenNLP com o módulo SentenceDetector treinado com o corpus Bosque[1] e a biblioteca Stanford CoreNLP[2] com o módulo LexicalizedParser, treinado com o corpus Cintil Treebank[3].

A etapa de **Identificação de Estrutura Independente** aplica uma nova segmentação para a árvore parse de cada sentença a fim de identificar estruturas independentes completas contendo uma estrutura menos complexa. Para isso utilizamos a biblioteca ExtroutMap[3] com o módulo OpenIE. Definimos por estrutura independente completa, aquela formada por sintagmas completos seguindo o padrão (1) e (2). Os sintagmas completos são: (i) sintagma NP, contendo um núcleo NN ou os seus derivados, (ii) sintagma VP, contendo um núcleo VB, ou derivado, e um sintagma completo NP, e (iii) sintagma PP, contendo um núcleo IN, ou derivado, e um sintagma completo NP. Estruturas intermediárias, sintagmas incompletos e tags existentes entre os sintagmas completos, são ignorados.

$$S < ((NP < (NN+)) \$ (VP < (VB+ \$ (NP < (NN+))) \quad (1)$$

$$S < ((NP < (NN+)) \$ (PP < (IN \$ (NP < (NN+))) \quad (2)$$

Na etapa de **Ajuste de Estrutura** são aplicados alguns ajustes sobre as estruturas independentes a fim de tornar mais compreensível e clara as futuras proposições. Para

[1] http://opennlp.apache.org
[2] https://stanfordnlp.github.io/CoreNLP
[3] http://extroutmap.lied.inf.ufes.br

isso, (i) regras morfológicas são aplicadas para identificar o núcleo dos sintagmas. Adotamos tokens {JJ, NN} para nomes e {VB, IN, RB} para relações; (ii) todos os *tokens* que pertencem ao núcleo NP são lematizados com o modulo Lemmatization da biblioteca CoreNLP; (iii) cada sintagma preposicional é transformado em um sintagma verbal por meio de um mapeamento; e (iv) relação de especialização é identificada por meio de nomes compostos e estrutura gramatical.

A etapa de **Extração de Triplas** objetiva extrair proposições que representem o fato expresso na estrutura independente completa. Para isso localizamos o primeiro sintagma verbal da estrutura e extraímos: sujeito, o sintagma nominal localizado antes do VP; objeto, o sintagma nominal localizado dentro do VP; e predicado, os *tokens* localizados entre o sujeito e objeto. A partir do sujeito, predicado e objeto é formada a proposição conceito-relação-conceito.

A etapa de **Ranking** é responsável por atribuir um peso para os conceitos seguindo algum parâmetro. Para isso, representamos a lista de conceitos na forma de um grafo, considerando que cada vértice possui um score hub, número de conexões de saída, e score authority, número de conexões de entrada. Assim, o peso *W* de cada conceito *k* é computado pela fórmula 3, cujo peso *W(k)* máximo é igual a 1.

$$W(k) = [\alpha . TF_d(k)] + [\gamma . (\rho . A(k) + \sigma . H(k))] \quad (3)$$

A fórmula associa o score hub *H(k)* e authority *A(k)* com a frequência dos conceitos no texto $TF_d(k)$. Os melhores parâmetros de ajuste do modelo HARD [Leake et al. 2004] foram atribuídos para $\rho$=2.235 e $\sigma$=1.764. Os parâmetros $\alpha$=0.3 e $\gamma$=0.7 foram adotados em fase de experimento.

A etapa de **Sumarização** é responsável por identificar as proposições relevantes diante do conjunto de triplas extraídas. Para isso aplicamos o conceito de quartis à topologia do grafo a fim de classificar os vértices, cujo peso de cada um é atribuído de acordo com a etapa de Ranking. Cada vértice é classificado como *heavy*, caso esteja localizado no terceiro quartil; *interjacent*, caso esteja localizado no caminho entre dois vértices heavy; *adjacent*, caso o peso do vértice seja superior ou igual ao menor peso dos vértices interjacent; e *light*, caso não se enquadre em nenhuma das classificações anteriores.

## 4. Experimentos e Resultados

Nesta seção apresentamos alguns resultados da abordagem proposta. Para realizar o experimento utilizamos como fonte de dados a seção Introdução do artigo [Novak e Canas 2008]. O texto está escrito no idioma português, sendo composto por 26 sentenças e 592 palavras. Foram realizados dois experimentos: (i) geração de mapa contendo todas as proposições extraídas da fonte de dados pela etapa de Extração de Triplas; e (ii) geração de mapa contendo as proposições filtradas pela etapa de Ranking e Sumarização.

### 4.1. Primeiro Experimento

O primeiro experimento identificou 26 sentenças, 123 proposições e 80 conceitos. A Figura 3 ilustra a saída deste processo sem aplicar as etapas de Ranking e Sumarização, ou seja, mostra todas as proposições identificadas a partir do texto.

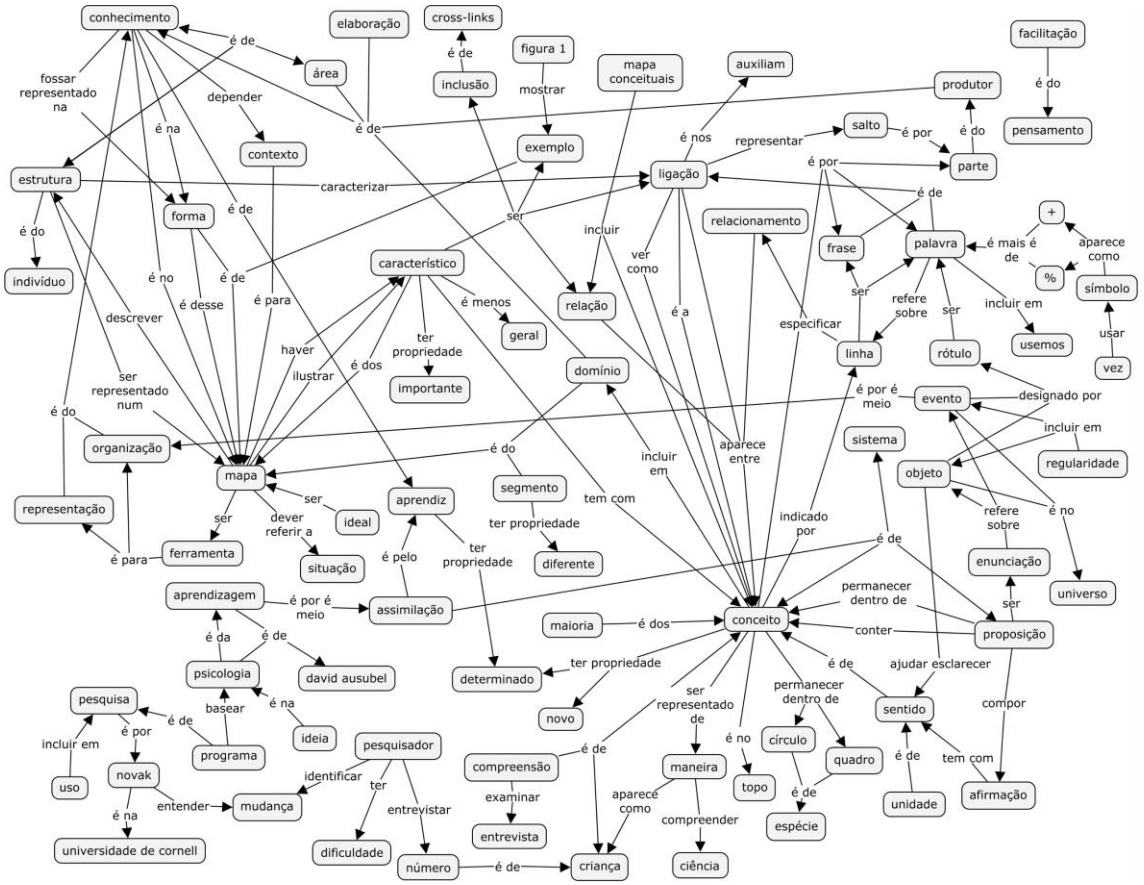



**Figura 3. Mapa conceitual extraído a partir do Primeiro Experimento**

A seguir apontamos algumas das características do mapa conceitual gerado pela abordagem:

- **Identificação de proposição a partir de sentença preposicional** – A proposição (relação, aparece entre, conceito) é extraída a partir do texto "...que são as relações ou ligações entre conceitos nos diferentes segmentos...". A abordagem cria uma relação com o rótulo "aparece entre" relacionando os conceitos "relação" e "conceito". Dado que consideramos mais expressiva e significativa aquela proposição formada por relação contendo verbo, definimos os rótulos das relações sempre contendo verbos.
- **Identificação de proposição a partir de relação de especialização** – Os conceitos "aprendiz" e "determinado" são extraídos a partir do texto "...conhecimento de um determinado aprendiz...". A abordagem criou uma relação de especialização entre os conceitos "aprendiz" e "determinado" com o rótulo "ter propriedade".
- **Resolução de Anáfora** – A proposição (psicologia, é de, david ausubel) foi extraída a partir do texto "...na psicologia cognitiva de Ausubel é que...". A abordagem associa o nome "Ausubel" para o conceito "David Ausubel" extraído a partir do texto "...baseava na psicologia da aprendizagem de David Ausubel...". Apenas anáforas contendo entidades nomeadas de pessoa foram resolvidas no experimento.
- **Identificação de Proposição a partir de conexões distantes** – Usando a árvore sintática criada a partir do texto "A Figura 1 mostra um exemplo de mapa

conceitual que descreve a estrutura dos mapas conceituais e ilustra as características acima.", a abordagem extrai as proposições distantes: (figura 1, mostrar, exemplo), (exemplo, é de, mapa), (mapa, descrever, estrutura) e (mapa, descrever, característica).

- **Identificação de conceitos *multi-words*** – A abordagem adota regras léxicas e sintáticas para identificar rótulos mais completos, tal como "mapas conceituais".

Conforme as características apresentadas a respeito do mapa conceitual construído pela abordagem, podemos perceber sua alta qualidade e diferença com os trabalhos relacionados. Porém, ainda temos alguns desafios: (i) A resolução de anáfora ainda está longe de ser satisfatória, especialmente em relação aos pronomes demonstrativos e possessivos. No experimento apenas anáfora de entidades nomeadas foram resolvidas; (ii) A qualidade e precisão do parser Português não é suficiente para as atividades linguísticas realizadas, sendo inferior a outros idiomas como o Inglês; e (iii) A lematização não atribuiu o lema correto em muitos conceitos, afetando o contexto e extração das proposições. Consequentemente, estes desafios em aberto causaram perda de informação e porção de mapa fragmentada.

### 4.2. Segundo Experimento

O segundo experimento adicionou a etapa de Ranking e Sumarização para o processo descrito na seção anterior, ou seja, aplicou o processo completo apresentado na Figura 2. O experimento identificou 53 conceitos relevantes e 95 proposições relacionadas. A Figura 4 ilustra a saída deste processo.

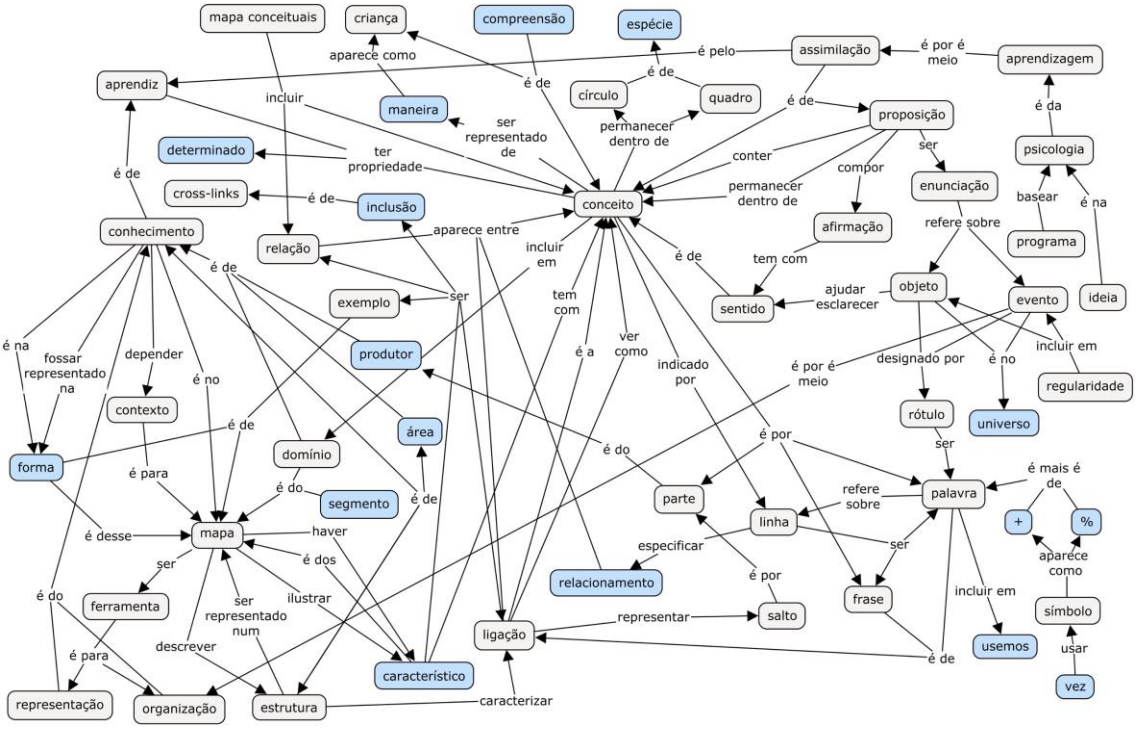



**Figura 4. Mapa conceitual extraído a partir do Segundo Experimento**

Uma análise foi conduzida subjetivamente comparando o mapa construído pela abordagem mostrado na Figura 4, com os mapas apresentados pelos trabalhos relacionados (Seção 2). Destacamos como principal contribuição da abordagem proposta, o uso do idioma Português, uma vez que os trabalhos relacionados usam idioma Inglês,

Croata e Espanhol. Além disso, notamos alguns pontos fortes associados com o mapa construído pela abordagem: (i) Todos os conceitos são conectados por frases de ligação, não havendo fragmentos ou conceitos livres; (ii) Rótulos são diretamente extraídos a partir da fonte de dados; (iii) Rótulo de conceitos são pequenos, não constituídos por pronomes e formados por *muti-words*, quando aplicável; (iv) Rótulo de relações são significantes e formados por verbos, algumas vezes não mencionado explicitamente no texto; e (v) Conceitos e proposições não exibem redundância. No entanto, notamos alguns pontos fracos, tal como: (i) Entidade nomeada são usadas como rótulos para conceitos; (ii) Informação importante do texto tem sido perdida; e (iii) Lematização de conceitos prejudicou a compreensão das proposições.

Para analisar a fidelidade do mapa conceitual com o texto, comparamos o mapa gerado automaticamente por nossa abordagem com mapas construídos manualmente por especialistas do domínio. Definimos por *especialistas*, estudantes de pós-graduação em informática na educação conhecedores de mapas conceituais. As seguintes instruções foram providas: (i) os especialistas receberam informações sobre o uso de mapas conceituais em geral e sobre o propósito do experimento; (ii) foram instruídos que tanto o rótulo de conceitos quanto de relacionamentos deveria ser curto, significativo e extraído a partir do texto; e (iii) foram informados que rótulos dos conceitos deveriam conter substantivos, e rótulos das relações deveriam conter verbos.

As tabelas a seguir mostram o *precision* e *recall* calculados comparando o mapa construído pela abordagem com os mapas construídos pelos especialistas. A Tabela 1 mostra a análise dos conceitos identificados, alcançando 0.68 em *precision* e 0.38 em *recall*. Neste experimento desconsideramos a flexão do rótulo dos conceitos, tal como plural.

**Tabela 1. Score da fidelidade dos Conceitos**

| **Especialista** | | **Esp.1** | **Esp.2** | **Esp.3** | **Esp.4** | **Esp.5** | **Média** |
|---|---|---|---|---|---|---|---|
| **Abordagem** | *Precision* | 0.77 | 0.69 | 0.65 | 0.70 | 0.59 | **0.68** |
| | *Recall* | 0.58 | 0.16 | 0.24 | 0.64 | 0.30 | **0.38** |

A Tabela 2 mostra a análise sobre as relações identificadas, obtendo 0.41 em *precision* e 0.19 em *recall*. Nesta avaliação, consideramos como relações similares aquelas que conectam os mesmos conceitos, exatamente, e que possuem significado aproximado.

**Tabela 2. Score da fidelidade das Relações**

| **Especialista** | | **Esp.1** | **Esp.2** | **Esp.3** | **Esp.4** | **Esp.5** | **Média** |
|---|---|---|---|---|---|---|---|
| **Abordagem** | *Precision* | 0.50 | 0.33 | 0.33 | 0.53 | 0.36 | **0.41** |
| | *Recall* | 0.29 | 0.05 | 0.08 | 0.41 | 0.11 | **0.19** |

O valor baixo alcançado pela métrica *recall* pode ser explicado devido o tamanho do mapa conceitual construído pelos especialistas. Uma vez que os especialistas estavam lendo o texto em seu idioma nativo e tinham domínio sobre o assunto, os mapas construídos foram muito breves e com mínimo conjunto de conceitos.

Além deste ponto, podemos destacar alguns outros relevantes que influenciaram o resultado: (i) Alguns rótulos atribuídos pela abordagem não corresponderam aos rótulos atribuídos pelos especialistas. A abordagem, às vezes, não faz uso de alguns adjetivos e advérbios relativamente importantes para caracterizar os conceitos; (ii) Alguns relacionamentos atribuídos pelos especialistas não foram explicitamente extraídos do

texto porque a informação pré-existente em sua estrutura cognitiva interferiu na construção do mapa; (iii) Alguns conceitos relevantes foram perdidos durante o segundo experimento devido a etapa de *ranking* e sumarização; e (iv) A atividade de extrair proposições a partir do texto e não a partir do conhecimento prévio do autor, requer muito tempo e grande esforço cognitivo, fato que prolongou a execução da atividade por mais de uma hora, afetando a qualidade dos mapas.

Portanto, para verificar a qualidade dos mapas conceituais construídos pelos especialistas, realizamos uma análise quantitativa comparando o mapa de cada especialista com todos os demais. A Tabela 3 mostra a análise sobre os conceitos identificados, alcançando score médio de 0.63 em *precision* e *recall*, inferior ao score obtido pela abordagem (Tabela 1).

**Tabela 3. Score dos Conceitos identificados pelos Especialistas**

| **Especialista** | | **Esp.1** | **Esp.2** | **Esp.3** | **Esp.4** | **Esp.5** | **Média** |
|---|---|---|---|---|---|---|---|
| **Esp.1** | *Precision* | 0.00 | 0.85 | 0.70 | 0.71 | 0.74 | **0.75** |
| | *Recall* | 0.00 | 0.28 | 0.35 | 0.85 | 0.50 | **0.49** |
| **Esp.2** | *Precision* | 0.28 | 0.00 | 0.45 | 0.27 | 0.44 | **0.36** |
| | *Recall* | 0.85 | 0.00 | 0.69 | 1.00 | 0.92 | **0.87** |
| **Esp.3** | *Precision* | 0.35 | 0.69 | 0.00 | 0.38 | 0.56 | **0.49** |
| | *Recall* | 0.70 | 0.45 | 0.00 | 0.90 | 0.75 | **0.70** |
| **Esp.4** | *Precision* | 0.85 | 1.00 | 0.90 | 0.00 | 0.81 | **0.89** |
| | *Recall* | 0.71 | 0.27 | 0.38 | 0.00 | 0.46 | **0.45** |
| **Esp.5** | *Precision* | 0.50 | 0.92 | 0.75 | 0.46 | 0.00 | **0.66** |
| | *Recall* | 0.74 | 0.44 | 0.56 | 0.81 | 0.00 | **0.64** |

A Tabela 4 mostra a analise sobre as relações identificadas, alcançando score médio de 0.42 em *precision* e *recall*, próximo ao obtido pela abordagem (Tabela 2).

**Tabela 4. Score da Relações identificadas pelos Especialistas**

| **Especialista** | | **Esp.1** | **Esp.2** | **Esp.3** | **Esp.4** | **Esp.5** | **Média** |
|---|---|---|---|---|---|---|---|
| **Esp.1** | *Precision* | 0.00 | 0.53 | 0.63 | 0.55 | 0.43 | **0.53** |
| | *Recall* | 0.00 | 0.14 | 0.27 | 0.71 | 0.23 | **0.34** |
| **Esp.2** | *Precision* | 0.14 | 0.00 | 0.33 | 0.11 | 0.30 | **0.22** |
| | *Recall* | 0.53 | 0.00 | 0.53 | 0.53 | 0.60 | **0.55** |
| **Esp.3** | *Precision* | 0.27 | 0.53 | 0.00 | 0.21 | 0.37 | **0.34** |
| | *Recall* | 0.63 | 0.33 | 0.00 | 0.63 | 0.46 | **0.51** |
| **Esp.4** | *Precision* | 0.71 | 0.53 | 0.63 | 0.00 | 0.57 | **0.61** |
| | *Recall* | 0.55 | 0.11 | 0.21 | 0.00 | 0.23 | **0.27** |
| **Esp.5** | *Precision* | 0.23 | 0.60 | 0.46 | 0.23 | 0.00 | **0.38** |
| | *Recall* | 0.43 | 0.30 | 0.37 | 0.57 | 0.00 | **0.42** |

Observamos que tanto o *precision* quanto o *recall* alcançados pela abordagem (Tabela 1 e Tabela 2) são próximos aos valores obtidos pelos mapas dos especialistas (Tabela 3 e Tabela 4). Por meio deste experimento, podemos observar que a atividade de construção de mapas a partir de textos é complexa e subjetiva, mesmo para especialistas, o que ressalta a dificuldade de construir automaticamente um mapa que represente o conhecimento consensual sobre o domínio de um texto.

Por fim, verificamos que embora o valor alcançado por nossa abordagem ainda não seja suficiente, apenas 16 conceitos (destacados em azul na Figura 4) dos 53 conceitos que compõem o mapa construído automaticamente não foram representados em um dos mapas construídos pelos especialistas.

## 5. Considerações Finais

O desenvolvimento de abordagens tecnológicas para construção automática de mapas conceituais a partir de textos tem mostrado resultados razoavelmente promissores. Para contribuir com os esforços em superar este desafio, apresentamos neste artigo uma nova abordagem baseada em técnicas linguísticas. A abordagem apresentada utiliza padrões gramaticais e busca em profundidade na árvore parser para identificação dos elementos do mapa conceitual. Nossa abordagem considera resolução de anáfora, mapeamento de preposições e identificação de multi-words. Para a relevância dos conceitos, a abordagem também apresenta um método combinando a frequência dos conceitos com a topologia do mapa; e para Sumarização apresenta uma classificação de vértices representados a partir do mapa.

Em geral, podemos afirmar, até agora, que a abordagem proposta apresentou resultados bastante aceitáveis, tanto quantitativos quanto qualitativos, para a identificação dos elementos constituintes de um mapa conceitual. Os experimentos demonstraram que há muitas dificuldades em construir manualmente um mapa a partir de texto, o que ressalta a importância de uma ferramenta computacional para auxiliar essa atividade. Trabalhos futuros estão direcionados à melhoria da abordagem proposta bem como ao desenvolvimento de uma biblioteca pública para extração de mapas. Ademais, os estudos englobam a melhoria de bibliotecas de processamento de linguagem natural no idioma português, cuja qualidade de processamento influencia diretamente no resultado da ferramenta proposta.

## Referências

Aguiar, C. Z., Cury, D., & Zouaq. (2015). A. Automatic Construction of Concept Maps from Texts. In Proceedings of the 7th Concept Mapping Conference.

Aguiar, C. Z., & Cury, D. (2016). A categorization of technological approaches to concept maps construction. In Learning Objects and Technology (LACLO). IEEE.

Leake, D., Maguitman, A., & Reichherzer, T. (2004). Understanding knowledge models: Modeling assessment of concept importance in concept maps. In Proceedings of the 26th conference CSS.

Novak, J. D., & Cañas, A. J. (2008). "The theory underlying concept maps and how to construct and use them.".

Villa, M., Aparicio, F., Maña, M. J., & de Buenaga, M. (2012). A learning support tool with clinical cases based on concept maps and medical entity recognition. In Proceedings ACM International conference on Intelligent User Interfaces.

Wang, W. M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2008). Mining knowledge from natural language texts using fuzzy associated concept mapping. Information Processing & Management, 44(5), 1707-1719.

Zouaq, A., & Nkambou, R. (2009). Evaluating the generation of domain ontologies in the knowledge puzzle project. IEEE Transactions on Knowledge and Data Engineering.

Žubrinić, K., Obradović, I., & Sjekavica, T. (2015). Implementation of method for generating concept map from unstructured text in the Croatian language. In Software, Telecommunications and Computer Networks (SoftCOM). IEEE.