Fusão Automática de Sentenças Similares em Português

Eloize Rossi Marques Seno, Maria das Graças Volpe Nunes

NILC – ICMC – Universidade de São Paulo
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

{eloize,gracan}@icmc.usp.br

Abstract. *This paper presents a Portuguese sentence fusion model. Sentence fusion is a text-to-text generation task which takes a set of similar sentences as input and combines these into a single output sentence. This process is of extreme relevance in many NLP applications, for instance, to treat redundancies in Multidocument Summarization by fusing information from a set of related sentences into a new one. We present three intrinsic evaluations of the model and the results obtained suggest that it has potential.*

Resumo. *Este artigo apresenta um modelo para a fusão automática de sentenças do Português. A fusão de sentenças é uma tarefa de geração de texto a partir de texto que, dado um conjunto de sentenças similares, produz uma nova sentença por meio da combinação de informações de várias sentenças do conjunto. Esse processo é desejável em várias aplicações do PLN, por exemplo, para eliminar informações redundantes na Sumarização Multidocumento, a partir da fusão de várias sentenças que expressam uma mesma informação em uma única sentença. Três avaliações intrínsecas do modelo são apresentadas e os resultados mostram que ele tem potencial.*

1. Introdução

Nos últimos anos há um crescente interesse por aplicações do Processamento de Língua Natural (PLN) que produzem textos a partir de textos (*text-to-text generation*, no inglês), em oposição à geração de textos tradicional baseada em uma representação não-lingüística subjacente à informação (como proposta por Reiter and Dale, 2000). Neste trabalho, o objeto de discussão é a fusão de sentenças, uma variante da geração de textos a partir de textos. A fusão de sentenças consiste em produzir, dadas duas ou mais sentenças similares de entrada, uma única sentença que combina informações daquelas sentenças, ao mesmo tempo em que elimina as informações redundantes.

Segundo Marsi and Krahmer (2005), a fusão sentencial pode ser de duas formas: por *interseção* e por *união* de informações. A fusão por *interseção* combina na sentença de saída somente as informações comuns que se repetem nas sentenças de entrada. A fusão por *união* preserva todas as informações das sentenças de entrada na sentença de saída. A Figura 1 apresenta um exemplo de *interseção* e de *união* de duas sentenças similares extraídas do corpus de trabalho (Seção 3). A escolha por uma forma ou por outra depende do objetivo da aplicação (Krahmer et al., 2008). Em Krahmer et al. (2008), experimentos realizados com usuários de um sistema de Perguntas e Respostas do domínio médico mostraram que a fusão por *união* é mais adequada nesse caso, pois há uma preferência maior por respostas mais longas. Já a fusão por *interseção* é de grande interesse na Sumarização Automática, especialmente na sumarização

multidocumento em que a redundância de informações é um problema (principalmente para os métodos extrativos), pois remete a um processo de síntese de sentenças com a preservação de conteúdo mais relevante (Barzilay and Mckeown, 2005). O modelo descrito neste artigo se baseia na interseção de informações redundantes em um conjunto de sentenças.

a) O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou a São Paulo às 18h45. b) A aeronave da TAM Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas.
Interseção: O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16. **União:** O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira com destino a Congonhas e chegou a São Paulo às 18h45.

Figura 1: Exemplo de fusão de sentenças por interseção e por união de informações

Nos trabalhos existentes para as línguas estrangeiras (por exemplo, Barzilay and Mckeown, 2005), a fusão de sentenças é composta de três passos: i) identificação de informações comuns, ii) fusão e iii) linearização. O primeiro passo consiste no alinhamento de informações similares, por exemplo, paráfrases e sinônimos entre as sentenças. Em geral, alinham-se árvores sintáticas (vide Pang et al., 2003) ou de dependências sintáticas (vide Barzilay and Mckeown, 2005 e Marsi and Krahmer, 2005). O segundo passo consiste na fusão de informações comuns previamente identificadas. Em algumas abordagens esse processo é realizado juntamente com o alinhamento. Em Pang et al. (2003), por exemplo, o alinhamento de árvores sintáticas é realizado de modo incremental, isto é, inicialmente um par de árvores é alinhado, obtendo-se uma floresta com as duas árvores e suas interseções. As sentenças restantes são alinhadas uma a uma com a floresta e unidas a ela. O resultado desse processo é uma floresta com a fusão de todas as sentenças de entrada. O último passo, por sua vez, consiste na escolha dos itens lexicais que irão compor a nova sentença e a realização da sentença em língua natural. A linearização envolve, portanto, os aspectos gramaticais da sentença a ser gerada. Na literatura, esse processo geralmente consiste em percorrer a árvore resultante da fusão gerando todas as sentenças possíveis, sem fazer uso de qualquer tipo de conhecimento. (vide Barzilay and Mckeown, 2005 e Marsi and Krahmer, 2005). Posteriormente, um modelo de língua é usado para selecionar a melhor sentença. Entretanto, como apontado nesses trabalhos, esses modelos não lidam adequadamente com as restrições de ordem das palavras, com os aspectos de concordância, entre outros. Assim, o modelo descrito neste artigo faz uso de conhecimento sintático obtido a partir das próprias sentenças de entrada e com o uso de um gerador de formas superficiais, tenta resolver esses aspectos de gramaticalidade, de modo a melhorar a qualidade das sentenças a serem selecionadas pelo modelo de língua.

O restante deste trabalho está organizado da seguinte da forma: a Seção 2 descreve o modelo de fusão de sentenças; a Seção 3 apresenta alguns experimentos realizados e, por fim, a Seção 4 apresenta as conclusões e possibilidades de trabalhos futuros.

2. Fusão de Sentenças

O modelo de fusão é composto por dois módulos principais: um módulo de alinhamento e fusão de informações comuns e um módulo de linearização (vide Figura 2). O sistema recebe

de entrada um conjunto de sentenças similares previamente processadas pelo *parser* Palavras (Bick, 2000), no qual cada sentença é mantida em um arquivo. Para cada sentença, o *parser* fornece informações de *part-of-speech* (*POS*) e de dependência sintática entre palavras e *chunks*, além do lema de cada palavra. Durante o alinhamento e fusão, o sistema faz uso da base de sinônimos Tep1 (Maziero et al., 2008), de uma *stoplist*, para a identificação de palavras irrelevantes ao alinhamento, e de um conjunto de regras que permitem o reconhecimento de paráfrases (Seção 2.1). Durante a linearização, um gerador de formas superficiais, desenvolvido no contexto do trabalho de Caseli (2007), é usado para auxiliar na realização da sentença (Seção 2.2). Como saída tem-se todas as fusões possíveis das sentenças de entrada. As subseções a seguir descrevem os dois módulos principais do sistema.

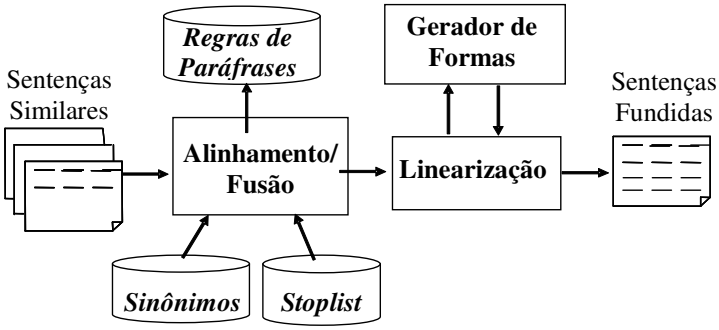

Figura 2: Ilustração do processo de fusão de sentenças similares

2.1 Alinhamento e Fusão de Informações Comuns

Dado um conjunto $S = \{s_1,...,s_n\}$ de sentenças similares, onde n >= 2, o algoritmo inicialmente identifica os segmentos (palavras ou *chunks*) semanticamente similares entre as duas primeiras sentenças s_1 e s_2. Para isso, ele se baseia em conhecimento lexical, sintático (i.e. informações de *POS* e traços de dependência), semântico (i.e. relações de sinonímia) e em regras de parafraseamento identificadas manualmente a partir de corpora, que permitem o reconhecimento de seqüências de palavras distintas, mas com o mesmo significado (por exemplo, *capital paulista* e *capital de São Paulo*). Cada segmento de s_1 é alinhado com no máximo um segmento de s_2. O alinhamento difere consideravelmente daquele realizado em outras tarefas do PLN (por exemplo, na Tradução Automática), pois alguns conceitos não têm correspondentes na outra sentença e, portanto, não são alinhados. Além disso, não há alinhamentos entre palavras de classes fechadas como artigos, preposições e conjunções. Essas palavras participam somente de alinhamentos envolvendo *chunks* (e.g. *a Mesa Diretora da Câmara* e *a direção da Câmara*). O algoritmo de alinhamento é descrito em detalhes em (Seno & Nunes, 2009).

Após alinhar s_1 e s_2, guarda-se para cada uma dessas sentenças todas as suas interseções com a outra sentença. Em seguida, o algoritmo tenta alinhar s_3 (se existir) com s_1 e s_2. O alinhamento de s_1 com s_3, por exemplo, é realizado de modo similar ao alinhamento de s_1 e s_2. O alinhamento de um segmento x pertencente a s_3 com um segmento y de s_1 que possui interseção com um segmento

[1] Disponível em: http://www.nilc.icmc.usp.br/tep2/download.htm (último acesso em 20/05/2009)

z de s_2, indica uma interseção também entre x e z. Assim, o algoritmo busca para cada segmento x_i de s_3 um segmento correspondente em s_1. Caso encontre, ele adiciona o segmento de s_1 como uma verbalização alternativa de x_i e também todas as interseções do segmento de s_1 com outras sentenças (se houver), terminando a busca por correspondentes de x_i. Do mesmo modo, x_i é adicionado como uma alternativa ao segmento de s_1. Caso não encontre nenhum correspondente em s_1, a busca continua em s_2. Esse processo segue até que todas as sentenças do conjunto tenham sido analisadas. A Figura 3 ilustra o alinhamento entre duas árvores de dependências sintáticas (Árvore 1 e Árvore 2), correspondentes às sentenças a) e b) (Figura 1), respectivamente. As setas indicam as dependências entre cada nó terminal e seu nó pai. Por exemplo, o nó terminal *Porto Alegre* (Árvores 1 e 2) é dependente do nó não terminal *partir* e representa uma relação de dependência entre verbo (ver) e objeto (obj). A árvore mais à direita na figura, ilustra a Árvore 1 após o alinhamento, na qual as caixas de textos e as setas não tracejadas representam as interseções com a Árvore 2, enquanto que as setas tracejadas indicam os nós sem alinhamento.

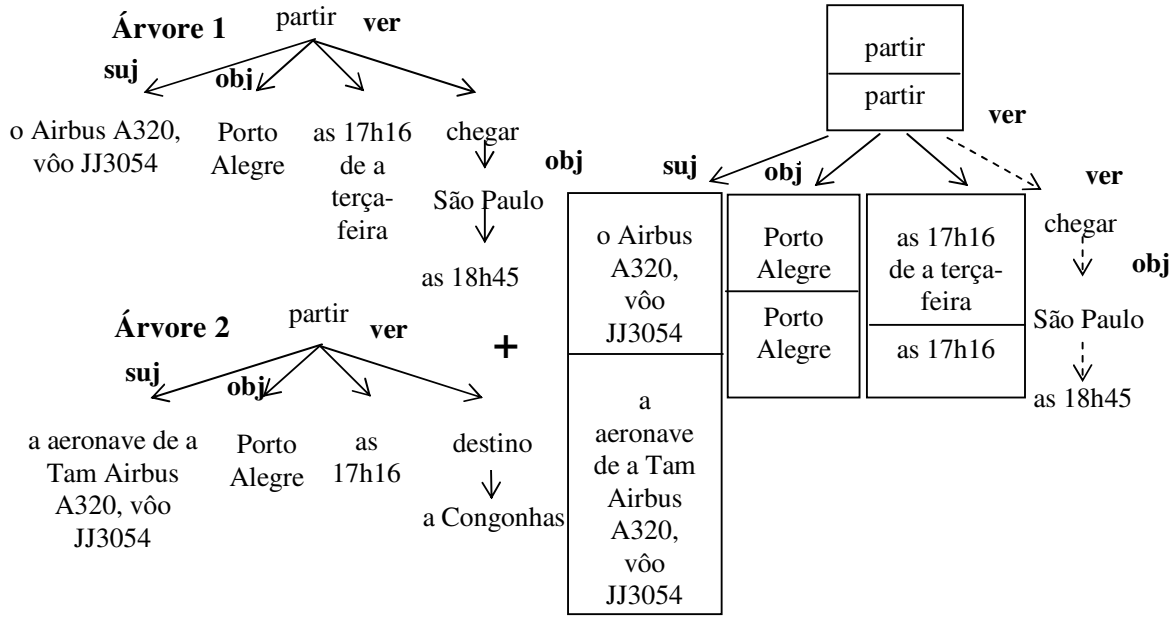

Figura 3: Exemplo de alinhamento de duas àrvores de dependências sintáticas

2.2 Linearização

A linearização consiste basicamente em escolher os itens lexicais entre as alternativas disponíveis, determinar a ordem de cada palavra na sentença e determinar a forma superficial mais adequada para cada palavra. Esse processo é baseado em conhecimento obtido das próprias sentenças de entrada. Mais especificamente, para cada sentença é mantido um vetor associativo com informações de todos os seus *tokens* (palavras e símbolos). As chaves correspondem à posição de cada *token* na sentença e guardam informações morfossintáticas, de *POS*, o lema, a relação de dependência (por exemplo, entre sujeito e verbo) e o seu dependente. Além disso, é mantido outro vetor com informações de cada *chunk* da sentença, por exemplo,

os sintagmas, as orações relativas e as orações adverbiais, sendo que para cada um deles são guardados os identificadores de todos os *tokens* que o compõem.

Dado que a fusão por interseção deve preservar somente as informações similares das sentenças de entrada, as informações que não têm qualquer interseção são removidas das sentenças. Entretanto, para evitar que constituintes sintáticos relevantes sejam excluídos, somente segmentos autocontidos que não interferem na gramaticalidade são removidos. Esses segmentos incluem as orações relativas, as orações adverbiais, uma oração em uma oração coordenada, entre outros. Na árvore da Figura 3, por exemplo, a subárvore não alinhada, que se refere a uma oração que é parte de uma oração coordenada (subárvore mais a direita), é excluída.

A geração da nova sentença envolve a seleção das palavras que melhor expressam um determinado conceito. Como não há informações semânticas suficientes para auxiliar essa decisão, o algoritmo gera todas as sentenças válidas possíveis a partir de cada árvore de dependência sintática contendo as interseções com outras sentenças, guiando-se por restrições de ordem das palavras herdadas das sentenças de entrada. Em alguns casos, as palavras alinhadas têm características morfossintáticas distintas e precisam ser modificadas, para não prejudicar a gramaticalidade. Por exemplo, considere que o sintagma nominal "*a parte concluída*" está alinhado ao sintagma "*o trecho*" e que "*a obra*" está alinhado com "*as reformas*" nas sentenças "*Na segunda etapa, a parte concluída será reaberta e a obra passará a ser feita na outra cabeceira.*" e "*Quando for concluída esta fase, o trecho será reaberto e as reformas passarão a ser feitas na outra cabeceira*.". Supondo que um percurso possível (válido) em uma das árvores resultasse em "*O trecho será reaberta e a obra passarão a ser feita na outra cabeceira.*", a sentença seria agramatical. Para previnir esses casos, o algoritmo faz uso de um gerador de formas superficiais (Caseli, 2007) que recebe como entrada o lema da palavra e suas características morfossintáticas e retorna a forma superficial correspondente. Assim, ele verifica, para cada sentença obtida, a concondância entre sujeito e verbo, verbo e objeto, substantivos e adjetivos, entre outros. Ao identificar, por exemplo, a discordância de gênero entre o substantivo "*trecho*" e o adjetivo "*reaberta*" (sentença anterior), o algoritmo obtém a forma superficial adequada para o adjetivo (i.e. "*reaberto*") com base nas características morfossintaticas do próprio substantivo. Já para obter a concordância entre o sujeito "*a obra*" e o verbo "*passarão*", ele busca na sentença original daquele sujeito as características morfossintáticas do seu pai, ou seja, do verbo "*passará*". As sentenças (a), (b) e (c) são exemplos de sentenças produzidas pelo sistema de fusão.

(a) O trecho será reaberto e a obra passará a ser feita na outra cabeceira.
(b) A parte concluída será reaberta e as reformas passarão a ser feitas na outra cabeceira.
(c) Quando for concluída esta fase, a parte concluída será reaberta e a obra passará a ser feita na outra cabeceira.

A seleção da melhor sentença gerada é feita com o auxílio de um modelo de língua baseado em bigramas. Esse modelo calcula a entropia de cada sentença de acordo com estatísticas derivadas de um corpus e a sentença com o menor valor de entropia é selecionada. O modelo foi induzido a partir do Corpus NILC[2], composto por 160 Mb de textos jornalísticos, usando o sistema jNina (Pereira and Paraboni, 2007).

[2] Disponível em: http://www.nilc.icmc.usp.br/~rh/corpus/ (último acesso em: 21/05/2009).

3. Experimentos

O modelo de fusão proposto foi avaliado em termos de seleção de conteúdo, gramaticalidade e semanticidade e de representatividade das sentenças produzidas. Para a avaliação, um corpus contendo 393 conjuntos de sentenças similares foi construído automaticamente, usando o algoritmo de clustering descrito em (Seno and Nunes, 2008). Os conjuntos foram obtidos a partir de 50 coleções de documentos comparáveis coletadas de várias agências de notícias da *web*. Dos 393 conjuntos, 100 foram selecionados aleatoriamente. A fim de eliminiar os grupos de sentenças irrelevantes da análise, os conjuntos com sentenças idênticas ou quase idênticas (com co-seno de similaridade maior do que 0.8) ou que não continham sentenças semanticamente similares foram excluídos (aproximadamente 43%).

A seleção de conteúdo foi avaliada comparando cada sentença gerada automaticamente com duas sentenças de referência produzidas por dois humanos. Para cada um dos 57 conjuntos (contendo de 2 a 4 sentenças cada), os humanos foram instruídos a produzir uma única sentença, preservando apenas as informações comuns entre elas. A concordância entre os humanos foi avaliada calculando a Precisão, a Cobertura e a *F-measure* de cada sentença do humano 1 (daqui a diante Ref1) em relação à sentença do humano 2 (Ref2). As sentenças foram segmentadas em nível de oração e para cada oração de uma sentença de Ref1 verificou-se se o seu significado havia sido preservado também na sentença de Ref2. Desse modo, a Precisão representa o número de orações de Ref1 preservado em Ref2 sobre o total de orações de Ref1. A Cobertura é dada pelo número de orações de Ref1 preservada em Ref2 sobre o total de orações de Ref2. Por fim, a *F-measure* representa a média harmônica entre a Precisão e a Cobertura. Os resultados obtidos, ou seja, 97% de Precisão, 89% de Cobertura e 91% de *F-measure* são próximos aos obtidos em trabalhos similares para a língua inglesa (i.e. 96% de *F-measure* (vide Barzilay and Mckeown, 2005)).

Os 57 conjuntos de sentenças foram processados pelo sistema e para cada conjunto foram selecionadas as 3 sentenças melhores pontuadas pelo jNina. As sentenças geradas iguais às sentenças de entrada foram previamente filtradas. Em dois casos, o sistema não gerou sentenças diferentes das de entrada e, portanto, nenhuma sentença pôde ser selecionada. A fusão, nesses casos, coincidiu com a menor sentença do conjunto. Em outro conjunto, somente uma das sentenças geradas diferia das sentenças de entrada, o que impossibilitou a seleção da segunda e da terceira melhor sentença. Em outros dois casos, apenas duas sentenças geradas não coincidiram com as sentenças originais, impossibilitando, assim, a análise da terceira melhor sentença desses conjuntos.

De maneira similar a comparação das sentenças produzidas pelos humanos, a Precisão, a Cobertura e a *F-measure* foram calculadas para as 3 melhores sentenças automáticas de cada conjunto (quando possível) em relação às sentenças de referência (vide Tabela 1).

Tabela 1: Resultados da avaliação de conteúdo para as 3 melhores sentenças

	(Ref1)			**(Ref2)**		
	Precisão	**Cobertura**	***F-measure***	**Precisão**	**Cobertura**	***F-measure***
Sent1	0,85	0,88	0,84	0,81	0,91	0,84
Sent2	0,84	0,94	0,87	0,81	0,97	0,86
Sent3	0,83	0,85	0,89	0,80	0,87	0,89

De acordo com a Tabela 1, o desempenho global do sistema tanto em relação à Ref1 como em relação à Ref2 foi praticamente o mesmo. Entretanto, ao analisar o desempenho em

relação a cada sentença, nota-se que a terceira melhor sentença apresentou o melhor resultado e a primeira melhor sentença obteve o pior desempenho (ainda não temos uma explicação para isso). Embora não seja possível fazer uma comparação direta com outros trabalhos da literatura, é válido dizer que em um experimento similar para a língua inglesa foi obtido um *F-measure* de 68% (vide Barzilay and Mckeown, 2005).

A avaliação da gramaticalidade, da semanticidade e da representatividade das sentenças geradas foi feita por outros dois humanos. Para gramaticalidade e semanticidade, os humanos foram instruídos a atribuir uma nota 3 para sentenças sem erros gramaticais ou semânticos; uma nota 2 para sentenças compreensíveis, mas com pequenos erros gramaticais ou semânticos; e uma nota 1 para sentenças sem sentido ou com sérios erros gramaticais. Os índices *kappa* obtidos em relação à avaliação da primeira, segunda e terceira melhor sentença foram 0,559, 0,446 e 0,582, respectivamente.

A avaliação da representatividade teve como objetivo analisar o quão bem cada sentença representa o seu conjunto, de modo que possa substituí-lo em um dado contexto sem comprometer significativamente sua mensagem principal. Os critérios para a representatividade foram: nota 3 para sentenças que expressam exatamente a mesma mensagem do conjunto; nota 2 para sentenças que expressam parcialmente a mensagem do conjunto; e nota 1 se a sentença distorcer a mensagem. Para essa avaliação, os índices *kappa* obtidos foram de 0,344, 0,322 e 0,303, para a primeira, segunda e terceira melhor sentença, respectivamente. Esses valores são considerados razoáveis, dada a natureza subjetiva da tarefa. A Tabela 2 apresenta os resultados médios obtidos.

Tabela 2: Resultados da avaliação de gramaticalidade, semanticidade e representatividade das 3 melhores sentenças geradas

	Gramaticalidade e Semanticidade		**Representatividade**	
	Humano1	**Humano2**	**Humano1**	**Humano2**
Sent1	2,4	2,6	2,7	2,6
Sent2	2,3	2,6	2,7	2,6
Sent3	2,2	2,4	2,6	2,6

Os resultados médios obtidos nas duas avaliações para a primeira, segunda e terceira sentenças são muito próximos, considerando ambas as referências. No que se refere à gramaticalidade e semanticidade, acredita-se que em alguns casos a avaliação pode ter sido prejudicada em função do modelo de língua ter atribuído uma pontuação melhor para sentenças com erros gramaticais e semânticos do que para sentenças corretas. Isso ocorre principalmente porque as sentenças mais curtas em geral têm um valor de entropia menor do que as sentenças mais longas, mesmo não sendo muito corretas. Apesar de não ser possível comparar diretamente esses resultados com os de outros trabalhos, vale dizer que em uma avaliação similar, porém considerando somente a gramaticalidade da melhor sentença gerada, o modelo de Barzilay and Mckeown (2005) obteve um valor de 2,3.

4. Conclusões e Trabalhos Futuros

Este artigo apresentou um modelo inédito para a fusão de sentenças em português que produz, a partir de um conjunto de sentenças similares, uma nova sentença, preservando as informações principais do conjunto. Em uma avaliação de desempenho, o sistema obteve 89% de *F-measure* no melhor caso, superando os resultados reportados em tarefa similar para a língua

inglesa (vide Seção 3). Experimentos também mostraram que, em muitos casos, as sentenças geradas são bem-formadas e preservam a mensagem principal do conjunto.

Os próximos passos incluem a extensão do modelo para permitir a fusão por união de informações, a aplicação dos modelos no contexto de Sumarização Multidocumento e a realização de experimentos avaliando a contribuição de cada modelo naquela tarefa.

Agradecimento

Agradecemos ao CNPq (Conselho Nacional de Pesquisa e Desenvolvimento) pelo suporte financeiro.

Referências

Barzilay, R, and McKeown, K. (2005). Sentence Fusion for Multi-document News Summarization, *Computational Linguistics*, Vol 31, nº 3, pp, 297-327.

Bick, E. (2000). *The Parsing System "Palavras" - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press.

Caseli, H.M. (2007). Indução de Léxicos Bilíngües e Regras para a Tradução Automática. Tese de Doutorado. ICMC-USP, 158 p.

Krahmer, E., Marsi, E. and van Pelt, P. (2008). Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion, In: *Proceedings of the Human Language Technology Conference – HLT*, pp, 193-196.

Marsi, E. and Krahmer, E. (2005). Explorations in Sentence Fusion. In: *Proceedings of the 10th European Workshop on Natural Language Generation – ENLG*, pp, 109-117.

Maziero, E.G., Pardo, T.A.S., Di Felippo, A., Dias-da-Silva, B.C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2,0 - Thesaurus Eletrônico para o Português do Brasil. In: *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana - TIL*, pp, 390-392.

Pang, B., Knight, K. and Marcu, D. (2003). Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In: *Proceedings of the Human Language Technology Conference – HLT/NAACL*, pp, 102-109.

Pereira, D.B. and Paraboni, I. (2007). A Language Modelling Tool for Statistical NLP. In: *Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, pp. 1679-1688.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge.

Seno, E.R.M. and Nunes, M.G.V. (2008). Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In: *Proceedings of the International Conference on Computational Processing of Portuguese Language - PROPOR* (Lecture Notes in Artificial Intelligence, 5190), pp, 133-144.

Seno, E.R.M. and Nunes, M.G.V. (2009). Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português. In: *Revista Linguamática*, Vol 1, pp.71-87.