

# Résumé multidocuments orienté par une requête complexe

Atefeh Farzindar<sup>1,2</sup>, Guy Lapalme<sup>2</sup>

<sup>1</sup>NLP Technologies Inc.  
farzindar@nlptechnologies.ca

<sup>2</sup>Université de Montréal, RALI  
lapalme@iro.umontreal.ca

## Résumé

Nous présentons un système de synthèse d'information pour la production de résumés multidocuments orientés par une requête complexe. Après une analyse du profil de l'utilisateur exprimé par des questions complexes, nous comparons la similarité entre les documents à résumer avec les questions à deux niveaux : global et détaillé. Cette étude démontre l'importance d'étudier pour une requête la pertinence d'une phrase à l'intérieur de la structure thématique du document. Cette méthodologie a été appliquée lors de notre participation à la campagne d'évaluation DUC 2005 où notre système a été classé parmi les meilleurs.

**Mots-clés** : synthèse d'information, résumés multidocuments, évaluation du résumé.

## Abstract

We present an information synthesis system for the production of multidocument summaries tailored by a complex query. We first analyze the user profile expressed by complex questions and we then compare the similarity between the documents and the questions at two levels : global and detailed. This research shows the importance of the study of the relevance of a sentence in the thematic structure of the document. This methodology has been applied for our participation in the DUC 2005 evaluation conference where our system was judged to be among of the best ones.

**Keywords**: information synthesis, multi-document summarization, evaluation of the summary.

## 1. Introduction

La synthèse d'information est définie comme le processus d'extraction, d'organisation et de correspondance entre des informations d'un ensemble de documents afin d'obtenir un rapport complet et non-redondant qui satisfait à un besoin précis d'information. Une forme particulière de synthèse d'information est un résumé multidocuments orienté par une requête. C'est un texte simple condensant un ensemble de documents avec une perte minimum d'information importante (Amigo *et al.*, 2004). À la différence des résumés indicatifs (qui aident à déterminer si un document est pertinent à une requête particulière), les résumés informatifs doivent répondre à des questions précises (Farzindar et Lapalme, 2003). Un autre exemple de synthèse d'information est le *Google Answer Service*<sup>1</sup> dans lequel l'utilisateur pose une question complexe à laquelle on ne peut répondre avec une recherche simple par un moteur de recherche. Google

---

<sup>1</sup> <http://answer.google.com>

demande ensuite de préciser le sujet et la catégorie de la question. Selon la complexité de la question, les détails demandés et le temps pour trouver la réponse, il faut attribuer une valeur entre 2\$ à 200\$ à payer pour que les experts de Google traitent la question.

La *Document Understanding Conference* (DUC) qui est une campagne d'évaluation annuelle portant sur le résumé automatique de textes. La tâche de la campagne 2005 avait changé considérablement par rapport à celles des années précédentes et comportait une notion de synthèse d'information. Les participants devaient produire des résumés d'une longueur n'excédant pas 250 mots à partir de groupe d'une trentaine de documents journalistiques et de questions complexes. Il y avait 50 groupes de questions pour lesquels il fallait produire un résumé. Voici un exemple d'un groupe de questions posées lors de cette évaluation :

How have people who have been accused of plagiarism reacted to the accusation? How has the accusation of plagiarism affected the person's success of reputation? Does the verity of accusation come into play?

De plus, chaque résumé doit répondre à deux granularités : générale et spécifique. La différence entre les deux granularités est la teneur en information spécifique, p.e. des cas particuliers, des dates, etc. Dans nos travaux précédents (Farzindar et Lapalme, 2005 ; Farzindar *et al.*, 2004), nous avons montré l'intérêt d'étudier de la pertinence d'une phrase dans son contexte thématique pour un domaine particulier comme le droit. Ce travail généralise cette idée à la synthèse d'information. Dans cet article, nous présentons CATS (*Cats is an Answering Text Summarizer*) (Farzindar *et al.*, 2005), le système que nous avons développé pour produire des résumés multidocuments orientés par une requête ainsi que les résultats obtenus lors de notre participation à l'évaluation de DUC 2005.

## 2. Synthèse d'information

Afin de produire un résumé cohérent à partir d'un ensemble des documents et d'une requête composée de plusieurs questions, nous analysons d'abord la requête pour dégager les informations recherchées. Ensuite, nous analysons des documents pour mettre en évidence les segments thématiques, contenant plusieurs phrases reliées autour d'un même thème, qui pourraient être pertinentes pour la question. Une fois identifiés les segments thématiques candidats pour chaque document, on construit une liste des phrases à partir des segments thématiques qui répondent mieux à la requête. Ce type de résumé, tenant compte d'un point de vue spécifique de l'utilisateur, demande une analyse détaillée des unités textuelles. Pour comparer les questions et les documents, nous avons d'abord analysé les entités nommées qui apportent les informations précises sur le sujet recherché dans la requête. Par la suite, nous découpons les questions et les phrases des documents en éléments de base (Hovy *et al.*, 2005). Le calcul de la similarité entre les éléments de base des questions et des phrases tient également compte des synonymes. L'élimination des informations redondantes et la compression des phrases sont importantes pour la sélection des unités afin d'intégrer le maximum d'informations différentes dans la limite des 250 mots.

## 3. Méthodologie

Dans notre approche, nous analysons le profil de l'utilisateur exprimé par des questions complexes en traitant chaque question et nous comparons les documents avec les questions à deux niveaux : global et détaillé. La comparaison globale fait ressortir les segments thématiques contenant les réponses aux questions. La comparaison détaillée effectue une analyse minutieuse

des segments thématiques choisis pour trouver les phrases pertinentes aux questions. Le résultat de l'évaluation montre que les résumés produits par notre approche sont plus cohérents car les phrases du résumé sont connectées par un thème.

### 3.1. Analyse des questions

Les questions contiennent : le sujet, les questions et la granularité du résumé. L'analyse des questions est effectuée en deux étapes : l'identification des types d'entités nommées et la décomposition des phrases en éléments de base.

#### 3.1.1. Entités nommées

Les entités nommées sont souvent une source d'information importante dans les résumés. Ceci est d'autant plus important lors de la génération d'un résumé de granularité spécifique qui devrait contenir plus de ce type d'information. Pour faciliter leur identification dans les textes sources, nous avons décidé d'identifier les catégories d'entités nommées qui doivent apparaître dans le résumé produit. Pour ce faire, nous utilisons donc certains concepts clés de la question. Pour CATS, nous avons considéré 4 catégories d'entités nommées : *person*, *organization*, *location* et *time*. Nous comptons le nombre de fois qu'un mot d'une catégorie se trouve dans la question, augmentant la probabilité que ce type d'entités nommées se trouve dans le résumé.

Par exemple, considérons la question suivante :

Identify and describe types of organized crime that crosses borders or involves more than one country. Name the countries involved. Also identify the perpetrators involved with each type of crime, including both individuals and organizations if possible.

*location*
*location*

*person*
*organization*

Dans cette question, nous pouvons identifier trois catégories : *location*, *person* et *organization*. Par conséquent, dans cet exemple, nous allons accorder plus de points aux phrases dans le texte source contenant des entités nommées de lieu, suivi de phrases contenant des entités nommées de personnes et d'organisations.

#### 3.1.2. Élément de base

Un élément de base (*Basic Element*) (Hovy *et al.*, 2005) est un triplet qui décrit la relation grammaticale entre deux mots dans une phrase. Il est composé des trois parties suivantes : l'entête, le modificateur et le type de relation. Puisqu'un élément de base ne varie pas de taille, il peut être facilement comparé à un autre élément de base permettant ainsi de mesurer leur similarité. Nous utilisons donc cette propriété pour comparer les éléments de base de la question et des phrases du corpus. Les éléments de base ont d'abord été introduits comme mesure d'évaluation de résumés dans le cadre de ROUGE (Lin, 2004) pour mesurer la similarité entre deux unités textuelles. L'originalité de CATS est d'utiliser également cette notion dans son traitement. Le module utilise l'analyseur syntaxique Minipar (Lin, 1998) pour construire l'arbre syntaxique qui est ensuite nettoyé et dont les relations entre les noeuds sont résolues pour produire une liste dont quelques exemples sont donnés dans le tableau suivant.

Entête	Mod.	Rel.
libyans	two	nn
indicted	libyans	obj
bombing	lockerbie	nn
indicted	bombing	for
bombing	1991	in

Pour CATS, nous éliminons certains éléments, qui ne nous sont pas utiles tels que les relations de type déterminant (*det*). Cette décomposition permet de faciliter la comparaison entre les phrases, parce qu'il y a moins de variations. Nous utilisons cette comparaison pour l'une des mesures de pointage des phrases.

### 3.2. Analyse des documents

#### 3.2.1. Comparaison globale

Pour la comparaison globale entre les documents et les questions, nous filtrons au niveau des segments thématiques. Nous utilisons la similarité du cosinus en calculant la ressemblance entre des vecteurs des poids des mots dans chaque segment thématique et la question pour ne garder que les segments qui nous intéressent. Ce filtrage est fait à partir d'un seuil minimal de similarité, choisi empiriquement. Pour la segmentation thématique, nous avons fait quelques expérimentations avec deux segmenteurs décrits par Hearst (1997) le système TEXTTILING et le segmenteur C99 décrit par Choi (2000) afin d'aligner les bornes de segment sur des frontières des découpages thématiques du texte. Les résultats obtenus par cet algorithme n'ont pas été concluants: trop souvent, l'algorithme ne séparait le texte qu'en deux ou trois segments thématiques, même pour des textes longs (1500 mots). Ces résultats sont nettement insuffisants, puisque nous voulions éliminer le plus de mots possible lors du premier filtrage pour le pointage des phrases. Il est à noter que l'unité de base de l'algorithme TEXTTILING est un paragraphe, contrairement à C99 qui utilise la phrase comme unité de base. Contrairement à l'algorithme C99, TEXTTILING produit beaucoup plus de segments thématiques: il n'est pas rare d'obtenir un segment par paragraphe. Malgré le fait qu'il produit autant de segments thématiques, nous avons décidé d'utiliser cet algorithme au lieu de C99 car nous voulions éliminer le maximum de mots lors du filtrage des segments. La multitude de segments nous permet donc de récupérer de façon plus concise l'information nécessaire à la production d'un résumé automatique.

#### 3.2.2. Comparaison détaillée

La comparaison détaillée entre les segments thématiques sélectionnés et les questions identifie les phrases les plus pertinentes correspondant à la requête. Afin d'identifier les mesures de pertinence des unités textuelles, nous avons entraîné notre système sur le corpus de DUC 2003 qui contient 30 ensembles de multi-documents et des questions de la collection TREC. Pour chaque phrase du segment thématique associé à la question, nous attribuons un pointage. À la fin les phrases sont triées en ordre décroissant de pointage qui est une combinaison linéaire des sept mesures suivantes:

**Éléments de base (Hovy *et al.*, 2005)** Nous comparons les éléments de base des phrases de la question avec ceux des phrases du corpus. Nous calculons un pointage selon la ressemblance entre les mots des constituants des éléments de base des deux phrases.

**Similarité du cosinus (Salton, 1989)** Nous appliquons directement le calcul de la similarité du cosinus entre les phrases de la question et les phrases du corpus de texte. Ceci permet de calculer la ressemblance entre deux vecteurs des poids des mots dans le document et la question.

**Poids de la phrase** La somme des poids de ses mots obtenus à partir de *TF.IDF*.

**Position absolue** Pointage selon sa position à l'intérieur de texte.

**Position relative** Pointage selon sa position à l'intérieur d'un paragraphe.

**Entités nommées** Nous incrémentons un compteur pour chaque entité nommée dans la phrase du corpus qui fait partie d'une des catégories d'entités de la question.

**Expressions prototypiques** Nous calculons le nombre d'expressions prototypiques des phrases indiquant une phrase ayant une plus forte probabilité de contenir des informations saillantes, par exemple une phrase de conclusion. Nous incrémentons un compteur pour chaque expression prototypique trouvée dans la phrase.

### 3.3. Posttraitement des phrases

Afin d'obtenir un résumé plus concis et cohérent, certaines opérations sont effectuées sur les phrases pour éliminer certaines parties de phrase moins importantes ou remplacer certaines expressions par d'autres plus concises. Les sections suivantes décrivent le traitement des phrases choisies.

#### 3.3.1. Résolution des expressions temporelles

Les textes journalistiques contiennent fréquemment des expressions temporelles relatives à la date de publication telle que *yesterday*, *2 days ago*, *last month*, etc. Lorsque les phrases sont incluses dans notre résumé, elles ont perdu leur référence temporelle et ces expressions n'ont plus de sens. Nous réglons ce problème en calculant, avec le module TEMPEX, la valeur absolue de ces expressions pour qu'elles aient un sens dans n'importe quel contexte. Le module utilise donc les informations temporelles contenues dans le texte, par exemple la date de publication du document, pour résoudre les expressions temporelles relatives. Pour ce faire, TEMPEX utilise une série d'expressions régulières pour, dans un premier temps, identifier ces expressions. Dans un second temps, TEMPEX ajoute des balises autour de ces expressions et indique la valeur absolue qu'il croit que l'expression vaut. Les balises utilisées sont du type *Timex2* (Gerber *et al.*, 2002). Voici une phrase sélectionnée pour la question concernant le plagiat:

A former editor for the Wall Street Journal sued the paper Tuesday for \$12.64 million, claiming that he was fired and his reputation smeared by a false charge of plagiarism.

Dans un résumé, *Tuesday* ne veut rien dire car on ne sait plus la date de l'article qui contenait cette phrase. Alors nous utilisons le module TEMPEX. La résolution de ces expressions s'effectue en deux étapes:

1. Les expressions temporelles sont identifiées par le module TEMPEX sur les documents d'origine.
2. Lorsque les phrases ont été choisies, nous remplaçons les balises TimeML pertinentes par un format de date approprié, soit *MM/JJ/AA* et nous ignorons les autres. Parmi les

expressions que nous laissons tomber, il y a des expressions comme *several days*, *weekly* ou simplement *1995*, qui n'a pas besoin d'être remplacé.

### 3.3.2. Élimination de redondance

Une des qualités des résumés est qu'ils ne doivent pas avoir d'information redondante. Ainsi, dans la listes des phrases candidates pour le résumé final, nous éliminons les phrases ayant trop d'information similaire avec les autres. Pour mesurer la quantité d'information, nous avons comparé la ressemblance entre les unités des phrases en tenant compte la similarité lexicale, les synonymes des unités textuelles et les informations exprimées par les entités nommées. Pour ce faire, nous utilisons de nouveau la similarité du cosinus pour identifier si un couple de phrases se ressemblent en fonction d'un seuil empirique. En plus de la similarité du cosinus, nous comparons les entités nommées des phrases, car il s'agit d'une bonne mesure pour savoir si deux phrases parlent du même sujet. S'il y a deux ou plus entités nommées identiques dans les deux phrases, nous considérons ces phrases comme similaires et ne nous gardons que celle qui a le plus haut pointage de la mesure de pertinence.

### 3.3.3. Compression des phrases

Les résumés demandés par DUC 2005 ne doivent pas dépasser 250 mots tout en donnant le plus d'information possible. Alors il est important d'éliminer les informations non essentielles à l'intérieur des phrases sélectionnées pour ainsi incorporer le plus de phrases différentes à l'intérieur du résumé. Tout d'abord, nous éliminons systématiquement le texte entre `()`, `[]`, `{}`, `-` et `–`. Pour les résumés qui demandent une granularité générale, nous utilisons l'analyseur syntaxique Collins (Collins, 1999) en collaboration avec le part-of-speech tagger TreeTagger (Schmid, 1994). L'analyseur syntaxique Minipar ne donne pas l'arbre syntaxique, il ne donne que des couples de mots et leur relation. Alors que analyseur de Collins donne l'arbre au complet et les subordonnés relatives (*SBAR*) qui nous intéresse. Nous parcourons l'arbre obtenu en éliminant toutes les branches qui correspondent à des subordonnés relatives commençant par les mots *who*, *when*, *where* et *which*. Cette opération permet d'enlever les descriptions de personne, lieu, etc. qui souvent ne sont pas nécessaires dans le résumé de granularité générale.

## 3.4. Sélection des phrases

Notre algorithme termine en choisissant parmi les phrases ayant obtenu les meilleurs pointages jusqu'à ce que le résumé contienne au plus 250 mots. Ensuite, nous ordonnons ces phrases par ordre croissant de date de son document original.

## 4. Évaluation et résultats

Avec le système CATS, nous avons participé à la campagne d'évaluation de DUC 2005. Nous présentons le résultat obtenu pour notre système parmi les 32 équipes participantes. L'évaluation des résumés se fait en trois étapes par NIST:

**Qualité** Une évaluation manuelle sur la qualité de la linguistique sur 5 points: **qualité grammaticale** le texte ne doit pas contenir des éléments non textuels ou des erreurs de ponctuations ou de casses des mots, **redondance** le texte ne doit pas contenir des informations redondantes, **clarté des références** les noms et les pronoms doivent être clairement référés

dans le résumé, **focus** les informations du résumé ne doivent pas déborder du sujet de la question, **cohérence et structure** le résumé doit avoir une bonne structure et les phrases doivent se suivre de façon cohérente. Les résultats de l'évaluation sont présentés dans le tableau 1. Globalement, nous nous classons 7<sup>e</sup> pour cette évaluation.

**Pertinence** Est-ce que le résumé répond bien à la question pour la granularité choisie ? Cette évaluation est surtout basée sur la quantité d'information que le résumé fourni pour répondre à la question. Les résultats sont présentés dans le tableau 2.

La pertinence d'un résumé est souvent considéré comme la plus importante évaluation lors de ce type de conférence. CATS se classe très bien (3<sup>e</sup> rang), à peine plus faible que le meilleur système.

	1	2	3	4	5	Moyenne
CATS	3.96	4.58	3.46	3.22	2.30	3.50
Meilleur système	4.06	4.48	4.16	3.92	3.22	3.97
Moyenne des systèmes	3.78	4.41	3.01	3.12	2.16	3.29
Moyenne des humains	4.81	4.90	4.94	4.89	4.77	4.86

Tableau 1. Résultats de la comparaison des 32 systèmes, pour les cinq questions concernant la qualité linguistique des résumés (évaluation sur 5, moyenne des 50 résumés). Le meilleur système représente le système ayant obtenu la meilleure moyenne générale. 1: Qualité grammaticale, 2: Redondance, 3: Clarté des références, 4: Focus, 5: Cohérence et structure.

	Total
CATS	2.72
Meilleur système	2.78
Moyenne des systèmes	2.39
Moyenne des humains	4.67

Tableau 2. Résultats pour la pertinence des résumés (évaluation sur 5, moyenne des 50 résumés)

**Évaluation automatique** ROUGE-1.5.5 est utilisé pour comparer les résumés automatiques avec des résumés produits à la main par la NIST. ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) est la mesure d'évaluation des résumés basée sur le calcul statistique de co-occurrence de n-grammes communs entre le résumé automatique est le résumé modèle. Par exemple, ROUGE-2 calcule le nombre de paires de mots successifs communs entre les résumés candidat et modèle. Seulement les pointages de rappel pour ROUGE-2 et ROUGE-SU4 sont utilisés dans le pointage officiel. Les résultats sont présentés dans le tableau 3. Étonnamment, l'évaluation automatique ne semble pas significative du tout: la variance des résultats est d'à peine 0.00008 pour ROUGE-2 et de 0.00024 pour ROUGE-SU4. Aucune conclusion ne peut être obtenue de ces résultats.

La figure 1 représente les résultats globaux pour tous les systèmes. Les résultats de l'évaluation montrent que notre approche de production du résumé multidocuments basée sur la comparaison

	Rouge2	RougeSU4
CATS	0.06	0.13
Meilleur système	0.07	0.13
Moyenne des systèmes	0.06	0.11
Moyenne des humains	0.10	0.16

Tableau 3. Résultats pour l'évaluation automatique avec ROUGE (pointage de rappel seulement, moyenne des 50 résumés)

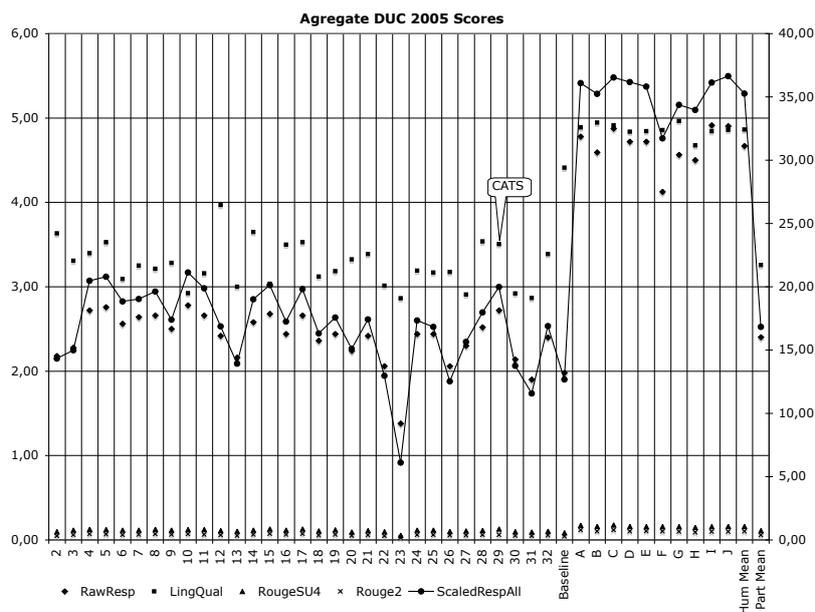


Figure 1. Graphique représentant les résultats globaux pour tous les systèmes. L'échelle de droite est pour le pointage de réceptivité des systèmes (ScaleRespAll) et celle de gauche pour tous les autres pointages. CATS se positionne parmi les 5 meilleurs systèmes sur l'ensemble des critères

globale au niveau la structure thématique du document et la comparaison détaillée au niveau des phrases avec les questions satisfait les critères d'évaluation. Pour les catégories les plus importantes comme la pertinence des phrases, la cohérence et les informations non-redondance notre système se positionne parmi les meilleurs systèmes participant à DUC 2005.

## 5. Conclusion

Dans cet article nous avons présenté notre approche de production du résumé multidocuments orienté par une requête complexe. Cette étude montre l'importance d'étudier de la pertinence d'une phrase pour une requête dans la structure thématique du document. Nous avons développé un système de synthèse d'information et nous l'avons évalué lors de la campagne d'évaluation de DUC 2005. La tâche définie pour cette compétition était de l'extraction des phrases pour former 50 résumés de 250 mots permettant de répondre à 50 questions complexes sur des sujets différents. Pour améliorer notre système et la concision des résumés produits, la compression des phrases candidates qui consiste à supprimer certains constituants syntaxiques des phrases et

régénérer le texte sans perte majeure d'information pour arriver à un résumé de 250 mots est un problème à considérer.

## Remerciements

Nous remercions particulièrement Frédéric Rozon qui a développé CATS et effectué les expérimentations au cours de l'été 2005. Nous remercions également Philippe Muller pour son aide dans le traitement des expressions temporelles. Cette recherche a été soutenue financièrement par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

## Références

- AMIGO E., GONZALO J., PEINADO V., PENAS A. & VERDEJO F. (2004). An empirical study of information synthesis tasks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Forum Convention Centre Barcelona.
- CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of NAACL-00*.
- COLLINS M. (1999). *Head-driven Statistical Models For Natural Language Parsing*. PhD thesis, University of Pennsylvania.
- FARZINDAR A. & LAPALME G. (2003). Using background information for multi-document summarization and summaries in response to a question. In *DUC03: NAACL'2003 Workshop in Automatic Text Summarization*, p. 168–173, Edmonton, Alberta, Canada.
- FARZINDAR A. & LAPALME G. (2005). Production automatique du résumé de textes juridiques: évaluation de qualité et d'acceptabilité. In *Traitement Automatique des Langues Naturelles (TALN 2005)*, p. 183–192, Dourdan, France.
- FARZINDAR A., LAPALME G. & DESCLÉS J.-P. (2004). Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte : solutions et perspectives*, 45(1), 39–65.
- FARZINDAR A., ROZON F. & LAPALME G. (2005). CATS a topic-oriented multi-document summarization system. In *DUC2005 Workshop*, p. 8 pages, Vancouver.
- GERBER L., FERRO L., MANI I., SUNDHEIM B., WILSON G. & KOZIEROK R. (2002). Annotating temporal information: From theory to practice. In *Proceedings of the 2002 Conference on Human Language Technology*, p. 226–230, San Diego, CA.
- HEARST M. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1), 33–64.
- HOVY E., LIN C.-Y., ZHOU L. & FUKUMOTO J. (2005). *Basic Elements*. <http://www.isi.edu/~cyl/BE/>.
- LIN C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain.
- LIN D. (1998). Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain, <http://www.cs.ualberta.ca/~lindek/minipar.htm>.
- SALTON G. (1989). *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc.
- SCHMID H. (1994). *TreeTagger - Decision Tree Tagger*. University of Stuttgart, <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTre%eTagger.html>.