Intégrer des ressources lexicales et grammaticales externes dans des analyseurs partiels probabilistes

Matthieu Constant[1] Isabelle Tellier[2]
(1) Université Paris-Est, LIGM, CNRS
(2) Université Paris 3, Lattice, CNRS
mconstan@univ-mlv.fr, isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Cet article traite de segmentation lexicale et de segmentation syntaxique. Nous montrons comment combiner des modèles probabilistes comme les champs markoviens aléatoires avec des ressources lexicales et grammaticales externes à l'aide d'une approche à états finis.

ABSTRACT

Integration of external linguistic resources in lexical and syntactic segmentation

This article focuses on lexical and syntactic segmentation. We show how to combine probabilistic models like conditional random fields with external lexical and grammatical resources using a finite-state approach.

MOTS-CLÉS : Segmentation lexicale, segmentation syntaxique, champs markoviens aléatoires, ressources lexicales et grammaticales.

KEYWORDS: Lexical segmentation, Syntactic segmentation, Conditional Random Fields, Lexical and grammatical resources.

1 Problématique

Dans cet article, nous nous focalisons sur la segmentation lexicale et syntaxique, qui correspond à l'étiquetage grammatical et l'analyse en constituants simples, incorporant la reconnaissance des mots composés. En particulier, nous décrivons une approche pour combiner des modèles probabilistes de segmentation et des ressources lexicales et grammaticales.

Une tâche de segmentation peut revenir à une tâche d'étiquetage simple en ajoutant des informations de segmentation dans les étiquettes (ex. BIO (Ramshaw et Marcus, 1995)). Par exemple, la phrase *Luc boit de l'eau de vie* peut s'étiqueter grammaticalement de la manière suivante *Paul/B-NPP boit/B-V de/B-DET l'/I-DET eau/B-NC de/I-NC vie/I-NC* où *de l'* est un déterminant composé partitif et *eau de vie* est un nom composé. Un analyseur probabiliste cherche à trouver la séquence d'étiquettes la plus probable étant donnée une séquence de même taille de mots (graphiques) en entrée. Les analyseurs *état-de-l'art* sont, en général, basés sur des modèles discriminants tels que maximum d'entropie ou les champs markoviens aléatoires (CRF).

Ces approches probabilistes dépendant uniquement d'un corpus d'apprentissage montrent cependant quelques limites. Par exemple, le traitement des mots simples et composés inconnus [1] cause de sérieuses chutes de performances des analyseurs. Des travaux récents se consacrent à l'utilisation de ressources lexicales riches comme sources de traits des modèles probabilistes afin de résorber en partie le problème : ex. (Denis et Sagot, 2009) pour un étiquetage simple, (Constant et Tellier, 2012) pour la reconnaissance des mots composés. C'est le genre de techniques que nous prolongeons ici.

2 Incorporation de ressources externes sous forme de traits

Nous utilisons les ressources externes comme sources de traits. Pour cela, nous nous basons sur une architecture à états finis (Blanc *et al.*, 2007). Celle-ci comporte d'abord une analyse ambigue du texte par consultation de dictionnaires électroniques et application d'une cascade de transducteurs. Cette procédure génère un automate acyclique représentant l'ensemble des analyses possibles. En particulier, l'analyse d'une séquence figée ou d'un constituant est représentée par une seule transition. Par exemple, l'analyse figée de *eau de vie* sera placée dans une seule transition étiquetée comme un nom. L'automate peut ensuite être élagué itérativement au moyen de différents modules de levée d'ambiguité basés sur des règles ou des heuristiques. A chacune de ces itérations, il est possible d'extraire des traits à partir de l'automate à chacune des positions dans le texte. Pour cela, on convertit l'automate au format BIO. Chaque transition correspondant à une séquence de m mots est convertie en une séquence de m transitions qui attribue une étiquette avec une information de type BIO à chacun des mots. Par exemple, la transition pour le mot composé *eau de vie* étiqueté comme un nom (NC) est décomposé en une séquence de 3 transitions : *eau/B-NC de/I-NC vie/I-NC*. La figure 1 donne l'automate issu de l'analyse lexicale dans le schéma BIO. Ainsi, à chaque position, on peut associer un ensemble de transitions duquel on peut extraire un certain nombre d'informations que l'on peut ajouter sous forme de traits au modèle. Par exemple, pour la position correspondant aux transitions sortant de l'état 5 (le mot *de*), on a l'ensemble d'étiquettes {I-NC,B-DET,B-PREP}. On peut en

1. qui ne se trouvent pas dans le corpus d'apprentissage.

particulier utiliser la concaténation des étiquettes de l'ensemble dans un certain ordre (ex. ordre alphabétique) comme trait.

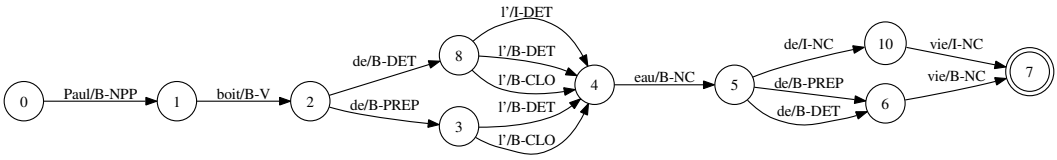

FIGURE 1 – Analyse lexicale dans le schéma BIO

3 Quelques expériences

Nous avons réalisé un certain nombre d'expériences sur le français, dans lesquelles notre corpus de référence était le corpus arboré de Paris 7 (Abeillé *et al.*, 2003), dans sa version de juin 2010. Ce corpus est composé d'articles journalistiques du Monde. Les mots composés sont marqués et considérés comme des unités lexicales. Nous avons utilisé les CRF (Lafferty *et al.*, 2001) et le logiciel *Wapiti* (Lavergne *et al.*, 2010) pour leur apprentissage et leur application. Les ressources lexicales et grammaticales étaient appliquées au moyen du logiciel *Unitex*[2].

Pour la segmentation lexicale, les ressources linguistiques externes sont composées de dictionnaires électroniques de mots simples et de mots composés, ainsi que de grammaires locales d'unités polylexicales semi-figées, compilées en transducteurs finis pour leur application. Nous avons observé des gains significatifs à la fois en terme de reconnaissance des mots composés (environ +5 points de F-mesure, pour des valeurs absolues autour de 70-80%) et d'étiquetage grammatical (+0.5 point, pour des valeurs d'absolues autour de 94%), par rapport au même segmenteur-étiqueteur sans ressources linguistiques externes.

Pour la segmentation syntaxique, nos ressources comprenaient, en plus, des grammaires locales de constituants syntaxiques simples. Nous avons montré, en particulier, leur grand intérêt sur des textes d'un autre domaine que le corpus de référence. Dans notre cas, nous avons travaillé sur un texte littéraire et des rapports parlementaires. Nous avons observé des gains de l'ordre de 8 points en terme de reconnaissance des constituants syntaxiques.

4 Conclusions

Ces expériences, comme celles de (Tellier et Dupont, 2013), montrent qu'il est avantageux de prendre en compte les informations véhiculées par un dispositif symbolique comme les automates à états finis dans des systèmes d'apprentissage automatique statistique comme les CRF. Les deux approches sont en effet complémentaires : les automates capturent des propriétés globales de la séquence, qui sont lexicalisées dans le codage BIO et injectées dans le CRF, qui à son tour évalue la confiance qu'il peut leur accorder pour le choix de l'étiquette finale.

2. http ://igm.univ-mlv.fr/˜unitex/.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Blanc, O., Constant, M. et Watrin, P. (2007). Segmentation in super-chunks with a finite-state approach. *In Proceedings of the Workshop on Finite-State Methods for Natural Language Processing (FSMNLP'07)*.

Constant, M. et Tellier, I. (2012). Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. *In Proceedings of the 8th conference on Language Resources and Evaluation (LREC'12)*.

Denis, P. et Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, pages 282–289.

Lavergne, T., Cappé, O. et Yvon, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 504–513.

Ramshaw, L. A. et Marcus, M. P. (1995). Text chunking using transformation-based learning. *In Proceedings of the 3rd Workshop on Very Large Corpora*, pages 88 – 94.

Tellier, I. et Dupont, Y. (2013). Apprentissage symbolique et statistique pour le chunking : comparaison et combinaisons. *In actes de TALN'2013*.