Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles CRF

Nicolas Pécheux[1, 2] Alexandre Allauzen[1,2] Thomas Lavergne[1, 2]
Guillaume Wisniewski[1, 2] François Yvon[2]
(1) Université Paris-Sud, 91 403 Orsay CEDEX
(2) LIMSI-CNRS, 91 403 Orsay CEDEX
{prenom.nom}@limsi.fr

Résumé. Quand on dispose de connaissances *a priori* sur les sorties possibles d'un problème d'étiquetage, il semble souhaitable d'inclure cette information lors de l'apprentissage pour simplifier la tâche de modélisation et accélérer les traitements. Pourtant, même lorsque ces contraintes sont correctes et utiles au décodage, leur utilisation lors de l'apprentissage peut dégrader sévèrement les performances. Dans cet article, nous étudions ce paradoxe et montrons que le manque de contraste induit par les connaissances entraîne une forme de sous-apprentissage qu'il est cependant possible de limiter.

Abstract.
Ignore what you know to better learn what you don't : a case study on type constraints for CRFs
When information about the possible outputs of a sequence labeling task is available, it may seem appropriate to include this knowledge into the system, so as to facilitate and speed-up learning and inference. However, we show in this paper that using such constraints at training time is likely to drastically reduce performance, even when they are both correct and useful at decoding. In this paper, we study this paradox and show that the lack of contrast induced by constraints leads to a form of under-fitting, that it is however possible to partially overcome.

Mots-clés : Étiquetage Morpho-Syntaxique ; Apprentissage Statistique ; Champs Markoviens Aléatoires.

Keywords: Part-of-Speech Tagging ; Statistical Machine Learning ; Conditional Random Fields.

1 Introduction

De nombreux problèmes de Traitement Automatique des Langues (TAL) peuvent être formalisés comme des problèmes d'étiquetage de séquences, bénéficiant de ce fait de méthodes et de résultats établis en apprentissage automatique. Il serait souhaitable pour certaines applications de pouvoir introduire des contraintes sur les étiquetages possibles, de manière implicite ou explicite afin d'introduire des connaissances linguistiques ou de réduire le temps de calcul pour des problèmes de grande dimension. Par exemple, dans une tâche de segmentation utilisant un encodage BIO [1] des étiquettes, on peut vouloir imposer qu'une étiquette 'O' ne précède jamais une étiquette 'I'. Les contraintes linguistiques peuvent introduire des connaissances provenant de règles syntaxiques ou de dictionnaires, comme c'est le cas dans certaines tâches d'analyse morpho-syntaxique (Li *et al.*, 2012). De manière plus pragmatique, l'analyse morpho-syntaxique pour les langues à morphologie riche implique de prédire une étiquette parmi des ensembles comprenant des centaines, voire des milliers, d'étiquettes : les problèmes de désambiguïsation associés sont donc à la fois plus difficiles et computationnellement plus coûteux, au point de rendre inopérantes les méthodes standard (Müller *et al.*, 2013).

Cette étude s'intéresse donc à l'introduction de contraintes lors de l'apprentissage d'un étiqueteur morpho-syntaxique. Pour cela, nous supposons disposer d'un dictionnaire associant à chaque mot un sous-ensemble des étiquettes possibles. Ce dictionnaire peut refléter une connaissance linguistique préalable, par exemple être extrait automatiquement de WIKITIONNAIRE ou encore être déduit des données d'apprentissage. Sous l'hypothèse que ce dictionnaire est correct, il semble naturel de vouloir prendre cette information en compte, afin d'une part, d'accélérer l'apprentissage et l'inférence, d'autre part, d'améliorer la qualité des prédictions. En réduisant l'ensemble des étiquettes pouvant être prédites et donc la taille

1. Pour début (B) ; intérieur (I) ; et en dehors (O).

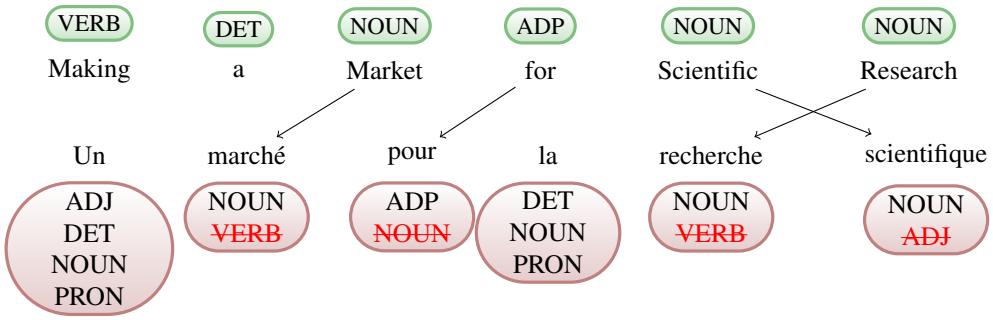

FIGURE 1: Instance d'apprentissage obtenue par transfert cross-lingue à partir d'une phrase source (haut) vers une phrase cible (bas). Les étiquettes autorisées par un dictionnaire sont représentées en rouge. Les étiquettes de la phrase source (en vert) sont *projetées* à travers les liens d'alignement de manière à désambiguïser les étiquettes cibles. Ces dernières constituent la référence (en noir), éventuellement ambiguë, pour le problème d'analyse morpho-syntaxique. À l'apprentissage, comme au décodage, on peut considérer, pour construire l'espace de recherche, que les 12 étiquettes sont possibles (cas non représenté) ou bien que seules celles qui sont proposées par le dictionnaire de type (encadrées en rouge) sont licites.

des espaces de recherche associés, les contraintes devraient permettre en un certain sens de simplifier la tâche du modèle. En effet, ce dernier peut alors concentrer son apprentissage sur la discrimination entre des hypothèses réalistes, en nombre réduit, plutôt que de considérer des configurations qui ne peuvent pas se produire.

Dans cet article, nous montrons que cette intuition n'est pas toujours correcte et qu'ajouter une telle information, même lorsqu'elle est pertinente et exacte, peut conduire à une dégradation de la capacité de généralisation du système, comme l'illustre le résultat paradoxal décrit à la section 2. La contribution de ce travail est d'apporter des explications à ce comportement inattendu, afin de pouvoir y remédier. Cette étude se concentre sur les modèles log-linéaires qui sont présentés à la section 3. En analysant théoriquement l'effet de l'inclusion des contraintes dans le modèle (section 4), il est possible de mettre en lumière les relations complexes qui existent entre les contraintes, la régularisation et le sous-apprentissage. Les résultats expérimentaux présentés à la section 5 montrent, en effet, que l'introduction de contraintes peut entraîner une forme de sous-apprentissage de certaines caractéristiques, qu'il est possible d'éviter. En particulier, il semble important de limiter, lors de l'apprentissage, l'impact des contraintes afin de garder une forme de contraste.

2 Un résultat paradoxal

Le point de départ de cette étude est une tentative de reproduire les résultats de Täckström *et al.* (2013). Ces derniers s'intéressent à une tâche d'analyse morpho-syntaxique pour des langues cibles peu dotées, pour lesquelles deux types de ressources sont disponibles : d'une part un dictionnaire (WIKTIONNAIRE) permettant de connaître, pour un mot, l'ensemble de ses catégories morpho-syntaxiques possibles ; d'autre part, un corpus parallèle aligné mot-à-mot et dont la partie source a été étiquetée automatiquement. En combinant ces deux ressources, comme illustré à la figure 1, il est possible d'apprendre un analyseur morpho-syntaxique, même lorsque l'on ne dispose pas de données cibles annotées.

Dans cette approche, le dictionnaire joue un rôle central : d'une part, en validant les étiquettes projetées au travers des liens d'alignement pour créer la référence ; d'autre part, en restreignant l'espace de recherche de l'analyseur morpho-syntaxique : lors de l'apprentissage et du décodage, la liste des étiquettes possibles pour chaque mot peut alors être réduite à un ensemble d'alternatives (les étiquettes autorisées par le dictionnaire) bien plus restreint que l'ensemble des étiquettes définies dans le schéma d'annotation.

Le tableau 1 rassemble les taux d'erreur obtenus par notre ré-implémentation du meilleur modèle décrit dans Täckström *et al.* (2013). En comparant la première et la deuxième ligne, on s'aperçoit qu'il est intéressant d'ajouter de manière explicite les contraintes de dictionnaire lors du décodage : ceci oblige le modèle à choisir, lors du décodage, l'une des étiquettes possibles et permet ainsi d'éviter certaines erreurs. Il est donc utile, du moins dans ce cas, d'utiliser l'information sur les étiquettes possibles d'un mot. En revanche, de manière surprenante, la troisième ligne de ce tableau montre qu'introduire ces contraintes lors de l'apprentissage dégrade sévèrement les performances.

Nous sommes donc, en apparence, face à un double paradoxe : (a) inclure des contraintes pourtant informatives pénalise le modèle ; (b) reproduire des conditions similaires à l'entraînement et au test n'est pas la meilleure configuration. Dans cet article, nous proposons d'expliquer ce paradoxe aussi bien d'un point de vue théorique (§ 4) qu'expérimental (§ 5).

appr.	test	cs	de	el	es	fi	fr	id	it	sv
✗	✗	17.3	13.3	16.8	14.7	19.2	14.1	14.8	13.3	12.5
✗	✓	16.7	11.8	16.3	12.4	17.4	13.7	14.6	12.7	12.0
✓	✓	21.2	15.8	17.6	15.5	27.4	23.1	27.9	15.1	14.7

TABLE 1: Une série de résultats surprenants : taux d'erreur (en %) pour neuf langues, obtenus par un modèle CRF partiellement observé sur une tâche d'analyse morpho-syntaxique par transfert cross-lingue à partir de l'anglais, selon que l'on utilise les contraintes de type pour définir l'espace de recherche à l'apprentissage (appr.) et/ou au test (test). L'intégration de contraintes à l'apprentissage dégrade systématiquement les performances. Les contraintes de type sont obtenues en prenant l'union d'un dictionnaire déduit des alignements et d'un dictionnaire extrait du WIKTIONNAIRE (voir la section 5.2 pour plus de détails sur les conditions expérimentales et les langues considérées).

3 Cadre classique : modèles et espaces de recherche

Cette section introduit un cadre général qui permettra de mieux comprendre le rôle des différents espaces de recherches manipulés dans notre problème d'apprentissage structuré ; la question des contraintes sera ensuite abordée au § 4.

3.1 Espaces de recherche et de référence

Dans un problème d'apprentissage générique, on dispose de l'ensemble $\mathcal{X}$ des *entrées* possibles, ainsi que celui $\mathcal{Y}$ des *sorties* possibles. Une manière classique de formuler un problème d'apprentissage pour le TAL (Smith, 2011, p. 23) est de considérer que pour chaque entrée $\mathbf{x} \in \mathcal{X}$, l'ensemble des sorties possibles est restreint à un sous-ensemble $\mathcal{Y}(\mathbf{x}) \subseteq \mathcal{Y}$. Cet espace $\mathcal{Y}(\mathbf{x})$ est appelé l'*espace de recherche*.

Exemple 3.1. Dans le cas de l'analyse morpho-syntaxique, notons $\mathcal{V}$ le vocabulaire et $\mathcal{T}$ l'ensemble des étiquettes morpho-syntaxiques possibles pour la tâche considérée. On a alors $\mathcal{X} = \bigcup_{n\in\mathbb{N}^*} \mathcal{V}^n$ et $\mathcal{Y} = \bigcup_{n\in\mathbb{N}^*} \mathcal{T}^n$. On considère usuellement, pour un $\mathbf{x} \in \mathcal{X}$ donné, uniquement les séquences d'étiquettes de même longueur que $\mathbf{x}$, c'est-à-dire que $\forall \mathbf{x} \in \mathcal{X}, \mathcal{Y}(\mathbf{x}) = \mathcal{T}^{|\mathbf{x}|}$.

Dans le cadre de l'apprentissage supervisé, on dispose d'un ensemble de données d'apprentissage $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1..N}$ supposées i.i.d sous une certaine distribution inconnue $\mathbb{D}$, où $\mathbf{y}_i \in \mathcal{Y}(\mathbf{x})$ est la sortie de référence pour l'entrée $\mathbf{x}_i$. De manière plus générale, on peut considérer que pour chaque exemple $\mathbf{x}_i$ on dispose d'un sous-ensemble $\mathcal{Y}^r(\mathbf{x}_i) \subset \mathcal{Y}(\mathbf{x}_i)$, que l'on appelle *espace de référence*. Tous les éléments de $\mathcal{Y}^r(\mathbf{x}_i)$ peuvent être complètement corrects, — ainsi, en traduction automatique, plusieurs traductions d'une même phrase peuvent être également bonnes —; ou bien seulement partiellement corrects comme dans le cadre de l'apprentissage partiellement supervisé, dans lequel on dispose d'une connaissance incomplète de la véritable référence.

3.2 Modèle log-linéaire

Dans ce travail, on s'intéresse aux modèles conditionnels log-linéaires qui, connaissant les entrées $\mathbf{x} \in \mathcal{X}$, définissent une distribution de probabilité sur les sorties $\mathbf{y} \in \mathcal{Y}(\mathbf{x})$ possibles comme :

$$p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})) \tag{1}$$

où $\phi(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ est un vecteur de d caractéristiques, $\boldsymbol{\theta} \in \bar{\mathbb{R}}^d$ un vecteur de paramètres et $\mathcal{Z}_{\boldsymbol{\theta}}(\mathbf{x})$ est le terme de normalisation. La log-vraisemblance des paramètres $\boldsymbol{\theta}$ s'écrit alors :

$$\ell_r(\boldsymbol{\theta}, \mathcal{D}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathcal{Y}^r(\mathbf{x}_i)|\mathbf{x}_i), \text{ avec} \tag{2}$$

$$p_{\boldsymbol{\theta}}(\mathcal{Y}^r(\mathbf{x})|\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{Y}^r(\mathbf{x})} p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}). \tag{3}$$

Remarquons que lorsque pour chaque $\mathbf{x} \in \mathcal{X}$, $|\mathcal{Y}^r(\mathbf{x})| = 1$, on retrouve le cadre classique de l'apprentissage supervisé. Le principe de maximum de vraisemblance consiste à choisir :

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ell_r(\boldsymbol{\theta}, \mathcal{D}) - \lambda_1 \|\boldsymbol{\theta}\|_1 - \frac{1}{2}\lambda_2 \|\boldsymbol{\theta}\|_2^2, \tag{4}$$

où l'on introduit souvent une régularisation $\mathcal{L}_1$ (pondérée ici par λ_1) et/ou une régularisation $\mathcal{L}_2$ (pondérée par λ_2). L'approche du maximum de vraisemblance conduit à augmenter la masse de probabilité de l'espace de référence $\mathcal{Y}^r(\mathbf{x})$ au sein de l'espace de recherche $\mathcal{Y}(\mathbf{x})$ (équation (3)). Dans le cas d'un CRF linéaire du premier ordre (Lafferty *et al.*, 2001), les caractéristiques se décomposent sur les paires d'étiquettes voisines selon :

$$\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \phi(y_i, y_{i-1}, \mathbf{x}). \tag{5}$$

La complexité de l'apprentissage et du décodage d'un CRF sont quadratiques en la taille du jeu d'étiquettes à l'ordre 1 et croissent exponentiellement avec l'ordre. Cette complexité justifie qu'on se limite en général à des jeux d'étiquettes restreints (typiquement quelques dizaines) et à des modèles d'ordre faible (1 ou 2).

4 Contraintes

Étant donnés le cadre de l'apprentissage structuré et les modèles log-linéaires décrits à la section 3, nous allons maintenant introduire formellement la notion de contrainte.

4.1 Fonction de contrainte et espaces restreints

Nous modélisons les contraintes par la notion de *fonction de contrainte* : $c : \mathbf{x} \in \mathcal{X} \to \mathcal{Y}^c(\mathbf{x}) \subseteq \mathcal{Y}(\mathbf{x})$. Ces fonctions sont déterministes et ne font pas partie du modèle.

Exemple 4.1. Dans le cas de l'analyse morpho-syntaxique, soit $t : \mathcal{V} \to 2^{\mathcal{T}}$ un dictionnaire associant à chaque mot un ensemble d'étiquettes, on considère la fonction « contrainte dictionnairique » suivante, que l'on note abusivement également t :

$$t : \mathbf{x} = (x_1, \ldots, x_{|\mathbf{x}|}) \in \mathcal{X} \to \mathcal{Y}^t(\mathbf{x}) = t(x_1) \times t(x_2) \times \cdots \times t(x_{|\mathbf{x}|})$$

qui n'autorise que les séquences d'étiquettes respectant, pour chaque mot, les contraintes données par le dictionnaire.

4.2 Modèle log-linéaire avec contraintes

On peut maintenant étendre les notations de la section 3 en prenant en compte des contraintes sur l'espace de recherche, données par une fonction de contrainte c :

$$\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}^c(\mathbf{x}), p^c_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}^c_{\boldsymbol{\theta}}(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})), \tag{6}$$

où le terme de normalisation devient :

$$\mathcal{Z}^c_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^c(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})). \tag{7}$$

Les contraintes influencent uniquement le calcul de la fonction de partition dans le modèle exponentiel : tout se passe comme si les sorties impossibles selon les contraintes avaient une probabilité nulle.

À l'apprentissage, en notant a la fonction de contrainte utilisée, l'équation (2) s'écrit :

$$\ell^a_r(\theta, \mathcal{D}) = \sum_{i=1}^{N} \log p^a_\theta(\mathcal{Y}^r(\mathbf{x}_i)|\mathbf{x}_i). \tag{8}$$

Tout comme à l'apprentissage, si l'on a accès à une fonction de contrainte de bonne qualité, il peut être avantageux d'exploiter celle-ci pour réduire les candidats possibles lors du décodage, et ainsi de diminuer les risques d'erreur tout en augmentant la vitesse d'inférence. En notant d cette fonction de contrainte, cela revient à considérer la règle de décision :

$$\mathbf{y}^* = f(\mathbf{x}) = \underset{\mathbf{y} \in \mathcal{Y}^d(\mathbf{x})}{\arg\max}\ p^d_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}). \tag{9}$$

Intuitivement, il semble préférable d'utiliser le même espace de recherche lors de l'apprentissage et du décodage, mais il est important de bien voir que rien ne l'impose. Nous avons d'ailleurs vu à la section 2 un exemple où il était préférable de ne pas considérer la même fonction de contrainte lors de l'apprentissage et lors du décodage (deuxième ligne du tableau 1). Remarquons que l'on pourrait même envisager de mettre des contraintes plus strictes à l'apprentissage que lors du décodage [2]. La question principale soulevée par cette étude est de se demander comment choisir optimalement $\mathcal{Y}^a(\mathbf{x})$ pour l'apprentissage et $\mathcal{Y}^d(\mathbf{x})$ lors du décodage.

4.3 Contraintes comme caractéristiques

Il est intéressant de noter qu'il est possible de représenter explicitement les contraintes, jusqu'ici externes au modèle, comme des caractéristiques particulières associées à des poids qui en dissuadent la violation. Soit c une fonction de contrainte et supposons qu'il existe un ensemble de caractéristiques $I \subset 2^d$ permettant d'encoder exactement le complémentaire de l'espace de recherche donné par l'application des contraintes :

$$\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}(\mathbf{x}) \qquad \mathbf{y} \notin \mathcal{Y}^c(\mathbf{x}) \Leftrightarrow \exists k \in I, \boldsymbol{\phi}_k(\mathbf{x}, \mathbf{y}) \neq 0.$$

Il suffit alors de fixer le poids de toutes ces caractéristiques à $-\infty$ pour obtenir un modèle sans contraintes équivalent. Par exemple, dans le cas des contraintes de type, on peut considérer l'ensemble des caractéristiques `(mot, étiquette)` qui ne figurent pas dans le dictionnaire [3]. On note que pour parvenir à cette représentation équivalente, il est nécessaire d'associer à ces caractéristiques un poids de $-\infty$, une valeur qu'il n'est pas possible d'atteindre dans la configuration sans contraintes du fait de la régularisation. L'utilisation explicite de contraintes revient donc, dans ce cas, à ignorer la régularisation pour une certaine classe de caractéristiques, ce qui peut donc conduire à du surapprentisage.

5 Expériences

Nous considérons dans cette section expérimentale deux tâches d'analyse morpho-syntaxique. En premier lieu, nous étudions en détail l'analyse morpho-syntaxique de l'allemand en considérant différents jeux de caractéristiques et d'étiquettes possibles, et en utilisant des contraintes extraites de différentes manières du corpus d'entraînement (section 5.1). Cette première tâche nous permet d'étudier les phénomènes de manière précise et contrôlée. En second lieu, nous nous intéressons à l'analyse morpho-syntaxique par transfert cross-lingue, dans laquelle les contraintes de type apparaissent naturellement. Dans cette seconde tâche, nous mesurons l'importance que peut prendre l'intégration des contraintes, si l'on souhaite obtenir des performances satisfaisantes.

5.1 Analyse morpho-syntaxique supervisée

5.1.1 Conditions expérimentales

Corpus et tâches On s'intéresse dans cette section à la tâche d'analyse morpho-syntaxique de l'allemand à partir de données annotées. Nous utilisons le corpus arboré TIGER (Brants *et al.*, 2004) avec le même partitionnement que Fraser *et al.* (2013), contenant 50 472 phrases, soit 888 238 mots étiquetés avec leur catégorie morpho-syntaxique. Les étiquettes pour cette tâche sont structurées en différents champs : la catégorie syntaxique (CS) pouvant prendre 54 valeurs possible, ainsi que des traits morphologiques (CM) : cas, nombre, genre, personne, temps, mode, pouvant prendre respectivement 4, 2, 3, 3, 2, 3 valeurs [4]. Ainsi, le mot *legendären* peut être étiqueté `cs=ADJ,cas=gen,num=sg,gen=masc,pers=✗,tmp=✗,mode=✗`. Sur les 1 373 étiquettes possibles, 619 sont observées sur le corpus d'apprentissage. Nous étudions les tâches consistant à prédire l'étiquette syntaxique (CS) et l'étiquette complète (CS+ CM).

2. Mais comme on peut s'y attendre, nous avons observé alors de très mauvaises performances, allant jusqu'à 90% d'erreurs.
3. Ces caractéristiques font typiquement partie des modèles, ce qui montre que ces derniers sont capable d'encoder implicitement les contraintes.
4. Ainsi que les valeurs « non applicable » et « ambigu » que l'on traite ici comme des catégories à part entière.

contraintes			MaxEnt				CRF d'ordre 1			
type	appr.	test	global	MDV	MHV	amb	global	MDV	MHV	amb
✗	✗	✗	10.7	6.6	49.8	10.9	2.9	2.2	9.4	3.0
corpus	✗	✓	10.7	6.6	49.8	10.9	2.9	2.3	8.8	3.0
	✓	✓	15.5	6.6	100.0	11.0	8.2	2.6	61.4	3.5
corr.	✗	✓	10.7	6.6	49.8	10.9	2.4	1.7	9.0	2.8
	✓	✓	15.4	6.5	100.0	11.0	7.7	2.1	61.6	3.4
oracle	✗	✓	6.1	6.6	1.0	10.9	1.6	1.7	0.4	2.8
	✓	✓	6.0	6.5	1.1	11.0	1.6	1.7	0.4	2.9

TABLE 2: Taux d'erreur (%) pour les modèles MaxEnt et CRF entraînés de façon supervisée pour la tâche d'analyse morpho-syntaxique sur le corpus TIGER, en fonction de différentes contraintes considérées à l'apprentissage (appr.) et/ou au test (test) : aucunes (✗) ; contraintes de type extraites du corpus d'apprentissage (corpus) ; corrigées en utilisant le corpus de test (corr.) ; et complétées (oracle). Le taux d'erreur est donné en prenant en compte tous les mots (global) ; les mots dans le vocabulaire (MDV) ; les mots hors vocabulaire (MHV) ; et les mots ambigus (amb).

Modèle Nous utilisons un CRF linéaire avec différents jeux de caractéristiques qui sont décrits par la suite. L'équation (4) est optimisée en utilisant 30 itérations de l'algorithme de propagation résiliente (Riedmiller & Braun, 1993). Nous utilisons une régularisation $\mathcal{L}_1$ et $\mathcal{L}_2$ dont les hyperparamètres sont choisis par *grid search*, pour chaque expérience, dans $\{0, 0.1, 1\}^2$ de manière à maximiser les performances sur le corpus de développement. Différents choix des contraintes de type impliquent un nombre très variable de caractéristiques et choisir la régularisation adaptée à chaque configuration est important pour ne pas interpréter à tort des différences de résultats qui seraient dues à une régularisation inappropriée.

Contraintes de type Nous envisageons trois manières différentes d'obtenir des contraintes de type à partir du corpus annoté : « *corpus* », « *corrigées* » et « *oracle* ». Les contraintes de *corpus* sont obtenues en considérant, pour chaque mot-type, l'ensemble des étiquettes auxquelles il est associé dans le corpus *d'apprentissage*. Par exemple « amüsiert » a pour seule étiquette ADJ dans le corpus. Cette méthode délivre cependant un dictionnaire incomplet (les mots hors du vocabulaire du corpus d'apprentissage ne sont pas couverts) et incorrect (certains mots ambigus ont pu n'être observés qu'avec une seule étiquette sur les données d'apprentissage). En effet, « amüsiert » apparaît également avec l'étiquette VERB en test. Afin d'étudier l'impact de ces deux problèmes sur les phénomènes étudiés, nous considérons deux conditions oracles, au sens où nous utilisons pour les définir les données de développement et de test. La première consiste à corriger le dictionnaire ainsi extrait : pour chaque mot dans le vocabulaire d'apprentissage, on s'assure que toutes les étiquettes observées en développement et en test sont bien incluses ; si ce n'est pas le cas, on les ajoute (contraintes « *corrigées* »). On associe donc à « amüsiert » les étiquettes ADJ et VERB. Dans la deuxième, on extrait les contraintes sur l'ensemble des données (de développement et de test) et non sur les seules données d'apprentissage (contraintes « *oracle* »). Cela revient également à considérer les contraintes corrigées auxquelles on a également ajouté les contraintes de type pour les mots hors du vocabulaire d'apprentissage.

Évaluation Les performances sont évaluées en utilisant le taux d'erreur standard (rapport du nombre d'occurrences incorrectes sur le nombre total d'occurrences) (global). Afin d'affiner davantage nos analyses, nous donnons aussi les taux d'erreur pour les mots connus (MDV) et pour les mots inconnus (MHV), et au sein de ces derniers, les taux pour les mots ambigus (c'est-à-dire observés dans le corpus d'apprentissage avec au moins deux étiquettes différentes) (amb).

5.1.2 Modèle MaxEnt simple

Nous commençons par un modèle log-linéaire très simple (MaxEnt) comprenant deux patrons de caractéristiques, le premier pouvant tester les associations `(mot, étiquette)` pour le mot et l'étiquette courante et le second testant l'étiquette courante seule `(étiquette)`. Notons que, dans ce modèle, il n'y a pas de dépendance entre étiquettes.

Les résultats obtenus par le modèle MaxEnt sont détaillés dans le tableau 2. Si, conformément à l'intuition (on prédit toujours l'étiquette la plus fréquente associée à un mot), ajouter les contraintes de type au test ne change rien aux résultats, on

peut s'étonner de l'impact négatif obtenu lorsque celles-ci sont incluses lors de l'apprentissage. Une analyse plus précise des résultats montre que cette baisse de performances est entièrement due aux mots inconnus : lorsque les contraintes oracles (qui incluent les étiquettes de tous les mots du corpus de test) sont considérées, les mots hors-vocabulaire sont systématiquement bien reconnus et les performances avec et sans contraintes au décodage sont équivalentes.

Cette expérience suggère que le principal problème lié à l'introduction des contraintes de type à l'apprentissage est de désambiguïser abusivement de trop nombreuses occurrences. En effet, une grande majorité des mots-formes du corpus d'apprentissage ne présentent pas d'ambiguïté et sont donc complètement désambiguïsés par les contraintes de type. Pour ces exemples, on a $p_\theta^a(\mathcal{Y}^r(\mathbf{x})|\mathbf{x}) = 1$ et l'occurrence ne contribue plus au gradient ni à l'optimisation de l'équation (4), ce qui implique qu'aucun paramètre n'est mis à jour. Dans le cas présent, cela entraîne en particulier que les paramètres relatifs aux *a priori* des catégories ne sont plus calculés que sur les mots ambigus. Or, les mots inconnus sont souvent plus proches des mots rares, eux-mêmes le plus souvent non-ambigus. Ainsi, l'étiquette associée à la caractéristique ayant le plus fort poids en l'absence de contraintes à l'apprentissage est NOUN, correspondant à 50% des mots inconnus, alors que l'étiquette ayant le plus grand *a priori* en appliquant les contraintes lors de l'apprentissage devient APPRART[5], qui ne correspond à aucun mot inconnu. On retrouve le fait que le filtrer des étiquettes équivaut à relâcher la régularisaion sur certaintes caractéristiques, ce qui peut conduire à sous-apprendre d'autres caractéristiques utiles, ici les caractéristiques relatives aux *a priori* des étiquettes.

5.1.3 Prédiction de la catégorie syntaxique avec un CRF

On considère maintenant un modèle CRF d'ordre 1 comportant un jeu de caractéristiques standard. Pour les mots courant, précédent et suivant, on considère : le mot en minuscule, ses préfixes jusqu'à une taille de 5, ses suffixes jusqu'à une taille de 2, s'il est en majuscule, s'il contient un trait d'union, s'il ne contient que des nombres, s'il contient un chiffre, sa forme obtenue en identifiant majuscules, minuscules et symboles, avec et sans répétitions (par exemple pour 'États-Unis' on a 'Xxxxx.Xxxx" et 'Xx.Xx). On considère également les associations des mots courant et précédent, des mots courant et suivant. Toutes ces caractéristiques sont considérées conjointement avec chaque étiquette possible, ce à quoi on ajoute l'étiquette courante seule et les bigrammes associant l'étiquette courante et les étiquettes suivante et précédente.

Les résultats obtenus par ce modèle sont dans le tableau 2 et sont au niveau de l'état de l'art (Müller *et al.*, 2013). Comme pour le modèle MaxEnt, les mots hors-vocabulaire constituent une part importante des erreurs. À nouveau, l'ajout des contraintes de type apprises sur le corpus d'apprentissage n'améliore pas les performances. En effet, comme illustré à la section 4.3, les caractéristiques `(mot, étiquette)` permettent d'apprendre les mêmes contraintes de manière endogène au modèle : on voit ici que cela est fait sans erreur. On observe, cependant, que corriger les contraintes issues de l'apprentissage permet d'obtenir des gains substantiels (réduction des erreurs de 2.9% à 2.4%). Cependant, que l'on corrige ou non les contraintes, les utiliser lors de l'apprentissage multiplie le taux d'erreur par un facteur d'environ trois. Le résultat paradoxal observé à la section 2 n'est donc pas spécifique au cadre du transfert cross-lingue ou de l'apprentissage partiellement supervisé. Ici encore, la dégradation observée pour les MHV explique une grande partie de la baisse des performances. On observe toutefois également une dégradation pour les mots ambigus présents dans le vocabulaire d'apprentissage. Contrairement à l'intuition initiale, réduire les candidats possibles pour permettre au modèle de n'avoir à discriminer que les étiquettes plausibles n'apporte, dans cette expérience du moins, aucun avantage. De manière intéressante, le phénomène disparaît dans la condition « oracle ». Comme le modèle est le même pour ces contraintes et pour les contraintes corrigées (puisque la seule différence est la prise en charge des MHV au test), on en conclut que savoir désambiguïser correctement les mots inconnus permet également de mieux prédire des mots voisins connus mais ambigus.

Deux hypothèses, mutuellement non-exclusives, peuvent expliquer ces résultats. La première, déjà évoquée à la section 5.1.2, met l'accent sur les mots complètement désambiguïsés par les contraintes de type à l'apprentissage. En effet, ces mots sont alors ignorés, alors que leurs statistiques et surtout les caractéristiques qu'ils partagent avec d'autres occurrences pourraient être utiles à d'autres endroits. Une seconde hypothèse est que l'introduction de contraintes rend les conditions d'apprentissage et de tests différentes, puisqu'à l'apprentissage tous les mots sont connus, ce qui n'est pas le cas au test. Cette incohérence entre l'apprentissage et test pourrait également contribuer à la dégradation des performances.

Pour tester ces deux hypothèses, nous avons effectué deux expériences de contrôle. La première essaie de résoudre le second problème en introduisant des mots inconnus lors de l'apprentissage. Les mots rares (c'est-à-dire de fréquence faible) ont souvent un comportement syntaxique proche des mots inconnus (Jurafsky & Martin, 2000, chap. 6). Nous proposons donc de ne pas utiliser les contraintes de type pour ces mots rares et de leur assigner, uniquement pour l'apprentissage, l'ensemble des étiquettes possibles. Dans nos expériences, nous considérons qu'un mot est rare si sa fréquence d'appa-

5. Préposition avec article.

contraintes	CS			CS + CM		
	global	MHV	amb	global	MHV	amb
✗	2.9	8.8	3.0	?	?	?
hapax10	3.0	9.3	3.1	?	?	?
hapax5	3.0	9.4	3.1	?	?	?
hapax1	3.2	11.2	3.1	14.4	37.2	14.0
min10	3.2	10.9	3.1	16.6	45.7	14.9
min4	3.3	12.8	3.0	17.5	53.4	15.2
min2	3.6	15.6	3.1	18.1	58.6	15.3
corpus	8.2	61.4	3.5	19.9	74.7	15.8

TABLE 3: Taux d'erreur (en %) d'un CRF supervisé pour la tâche d'analyse morpho-syntaxique sur le corpus TIGER avec le jeu d'étiquettes syntaxiques seules (CS) ou syntaxiques et morpho-syntaxiques (CS + CM), en considérant à l'apprentissage les contraintes de type : uniquement pour les mots de fréquence supérieur à 10 (hapax10), 5 (hapax5), 1 (hapax1) ; en s'assurant que toute position comprend un minimum d'étiquettes (min10, min4, min2) ; ou pour tous (corpus). Lors du test on utilise systématiquement les contraintes *corpus*. Pour la tâche d'analyse morpho-syntaxique complète, il n'est, dans certains cas, pas possible de faire l'expérience en un temps raisonnable.

rition est inférieure à un (hapax1), à cinq (hapax5) ou à dix (hapax10). Le tableau 3 montre que cette heuristique permet partiellement de résoudre le problème observé. Pour le modèle simple (MaxEnt), cette heuristique suffit à ramener le modèle avec contraintes de type à l'apprentissage au même niveau que le modèle sans contrainte. Pour les modèles CRF, plus riches en caractéristiques, la dégradation est faible dans le cas des conditions hapax5 et hapax10. On observe encore une fois que l'amélioration des performances résulte principalement d'un meilleur traitement des mots inconnus.

Le problème de l'approche précédente est que pour les mots rares, toutes les étiquettes sont considérées, ce qui reste problématique dans des tâches où cet ensemble est très grand. Comme la difficulté semble surtout provenir des mots complètement désambiguïsés à l'apprentissage, dans une deuxième expérience, nous ajoutons pour chaque mot complètement désambiguïsé un certain nombre d'étiquettes aléatoires de manière à s'assurer qu'il y a au moins i compétiteurs (min-i) au total (en comptant l'étiquette de référence) à chaque position. Les résultats présentés dans le tableau 3 montrent que les performances obtenues sont bien meilleures que lorsque l'on applique les contraintes de base, et légèrement moins bonnes que lorsque l'on autorise toutes les étiquettes pour les mots rares. Il est enfin possible de n'ajouter des étiquettes aléatoires que pour les mots *rares* (et désambiguïsés par les contraintes), mais il s'avère que cela détériore légèrement les résultats. Il semble donc que le nombre de concurrents joue également un rôle important pour les performances.

5.1.4 Analyse morpho-syntaxique pour le jeu d'étiquette complet

Nous considérons ensuite la tâche de prédiction de l'étiquette morpho-syntaxique complète. Les étiquettes morpho-syntaxiques sont structurées, au sens où les catégories morphologiques possibles dépendent de la catégorie syntaxique. Une approche possible est donc de découpler le problème en apprenant d'abord les étiquettes syntaxiques, puis en utilisant celles-ci pour filtrer les traits morphologiques (Müller *et al.*, 2013). Dans ce travail, nous nous intéressons toutefois uniquement à la prédiction de l'étiquette morpho-syntaxique complète.

Pour cette tâche, le nombre total d'étiquettes rend prohibitive l'utilisation du modèle CRF précédent, en l'état, et diverses heuristiques doivent être envisagées pour réduire l'espace de recherche. Il est, de plus, nécessaire de limiter le nombre d'étiquettes candidates pour les mots inconnus. Une première approche consiste à se limiter à l'ensemble des étiquettes observées (619 au lieu de 1373), ou bien encore aux étiquettes dites « ouvertes » [6], ce qui limite les étiquettes possibles à 435, ou enfin, selon l'approche retenue dans cet article, de ne prendre en compte que celles qui sont observées avec des mots rares (de fréquence 1), ce qui ramène ce nombre à 204. Nous considérons les mêmes caractéristiques que pour le modèle de la section 5.1.3, à ceci près que chaque fois que l'on considérait une caractéristique portant sur une étiquette (catégorie syntaxique), nous considérons maintenant à la fois l'étiquette complète, la catégorie syntaxique et les combinaisons impliquant la catégorie syntaxique et chacune des catégories morphologiques. On aura donc, par exemple, une

6. Estimées en partitionnant les données d'apprentissage et en imposant que la fréquence à laquelle une étiquette est vue avec un nouveau mot soit supérieure à un seuil (e.g. 10^{-4}).

caractéristique testant à la fois la catégorie syntaxique et le cas des étiquettes courante et précédente. À notre connaissance, seuls Müller *et al.* (2013) et Silfverberg *et al.* (2014) ont également utilisé des caractéristiques internes aux étiquettes.

En utilisant les contraintes de type, il est possible d'entraîner un modèle CRF standard sans avoir besoin d'utiliser d'autre heuristique simplificatrice (contrairement, par exemple, à Müller *et al.* (2013)). Les résultats obtenus, dans le tableau 3, montrent que l'on retrouve le même comportement que pour la tâche d'analyse syntaxique simple. Garantir un minimum de compétiteurs à chaque position permet un bon compromis entre vitesse d'apprentissage et performances en généralisation. Ceci est insuffisant cependant pour obtenir les mêmes performances qu'en omettant les contraintes pour les mots rares (hapax1), résultat alors état de l'art pour un modèle d'ordre 1 (Müller *et al.*, 2013), mais un peu plus de dix fois plus lent à entraîner. On peut imaginer augmenter encore les performances au prix d'un entraînement plus long. De meilleures techniques qui permettraient de limiter les dégradations dûes à l'introduction de contraintes de type, tout en conservant leur bénéfice computationnel, restent à trouver.

5.2 Analyse morpho-syntaxique ambiguë

La tâche d'analyse morpho-syntaxique de la section 5.1 nous a permis d'étudier le problème dans un cadre bien contrôlé. Dans ce cadre, les contraintes de type, même lorsqu'elles ne sont utilisées qu'au décodage, ne permettent jamais d'améliorer les performances. Il s'avère en fait que le modèle est capable de les apprendre presque parfaitement, et leur seul intérêt provient du gain important en vitesse qu'elles permettent. Ces contraintes étant exclusivement extraites du corpus d'apprentissage lui-même, elles n'apportent toutefois aucune information nouvelle, et peuvent donc être responsables du surapprentissage observé. Nous considérons ici un autre exemple, dans lequel les contraintes de type apparaissent de manière naturelle, sont extraites indépendamment du corpus d'apprentissage et se révèlent utiles pour améliorer les performances.

On considère la tâche d'analyse morpho-syntaxique faiblement supervisée, introduite à la section 2 et décrite en détail dans (Täckström *et al.*, 2013). Nous reproduisons la configuration de Wisniewski *et al.* (2014) en utilisant leurs ressources ainsi que le code fourni [7]. Pour chaque langue, Wisniewski *et al.* (2014) utilisent deux sources de contraintes de type : d'une part un dictionnaire automatiquement extrait de WIKTIONNAIRE et d'autre part des contraintes extraites du corpus d'apprentissage, annoté indirectement à partir de l'anglais à travers des liens d'alignement. Les contraintes extraites des bitextes jouent un rôle analogue aux contraintes extraites des corpus de la section 5.1, alors que les contraintes issues du WIKTIONNAIRE reflètent une connaissance linguistique externe que l'on souhaiterait exploiter. En plus d'être utilisées pour apprendre la référence ambiguë, c'est-à-dire pour construire $\mathcal{Y}^r(\mathbf{x})$, les contraintes de type c peuvent restreindre l'espace de recherche $\mathcal{Y}^c(\mathbf{x})$ à l'apprentissage et au décodage, configuration retenue par Täckström *et al.* (2013) et Wisniewski *et al.* (2014). Les tableaux 1 et 4 montrent pourtant que comme pour l'apprentissage supervisé, inclure les contraintes à l'apprentissage nuit aux performances, avec dans certains cas une différence drastique, par exemple pour le finnois (fi) ou l'indonésien (id) pour lesquels le taux d'erreur est quasiment doublé. En omettant les contraintes de type lors de l'apprentissage, ce simple changement permet d'obtenir un gain moyen de 6.0% sur les langues considérées par rapport au modèle [8] état-de-l'art de Täckström *et al.* (2013), ce qui montre l'importance de prendre en compte le problème.

La section 5.1 permet de comprendre en quoi les contraintes de type posent problème lors de l'apprentissage. En effet, pour toutes les positions sans lien d'alignement, les étiquettes de référence sont les mêmes que les étiquettes possibles : $\mathcal{Y}^r(\mathbf{x}) = \mathcal{Y}^a(\mathbf{x})$ et donc $p^a_\theta(\mathcal{Y}^r(\mathbf{x})|\mathbf{x}) = 1$ (voir l'équation (8)). Le modèle n'apprend donc rien pour cette position. Dans la figure 1, par exemple, le premier mot 'Un' est associé à quatre étiquettes références possibles, qui sont aussi les étiquettes possibles lorsque les contraintes de type sont appliquées à l'apprentissage. Cette observation est toujours vraie si les contraintes de type permettent de désambiguïser complètement un mot. Utiliser les contraintes de type revient donc à ignorer une grande partie des exemples. Une solution possible est alors d'utiliser à l'apprentissage les contraintes de type uniquement si les étiquettes ainsi restreintes sont strictement plus nombreuses que les étiquettes de référence. Par exemple, dans la figure 1, pour la première position, 'Un' possède quatre étiquettes références possibles ; on utilise donc l'ensemble des douze étiquettes possibles pour définir l'espace de recherche. En revanche, pour la deuxième position, 'marché', seule

7. `http://perso.limsi.fr/wisniews/ambiguous`

8. En réalité, à cause d'une erreur d'implémentation, les résultats publiés dans Täckström *et al.* (2013) correspondent au cas où l'on applique aucune contrainte de type pour réduire l'espace de recherche (premières lignes de chaque block dans le tableau 4). Les résultats corrigés ont été publiés par la suite dans un errata disponible ici `http://www.dipanjandas.com/files/erratum.pdf`. Bien que les auteurs considèrent que les résultats des deux configurations sont semblables, leurs résultats montrent pourtant clairement une différence de l'ordre de 2% en moyenne pour le cas des contraintes bitexte seules et de l'union, mais pas dans le cas des contraintes issues de WIKTIONNAIRE seules. Nous observons dans nos expériences une différence pour ce dernier cas également, que nous attribuons à la manière dont sont extraits les dictionnaires par Wisniewski *et al.* (2014) qui utilisent toutes les informations de forme de WIKTIONNAIRE, et donc obtiennent des contraintes de type plus puissantes (et donc plus dangereuses à l'apprentissage).

contraintes	appr.	test	cs	de	el	es	fi	fr	id	it	sv
bitexte	✗	✗	17.3	13.6	17.0	14.8	19.2	14.3	14.8	13.5	12.4
	✗	✓	17.3	12.3	17.5	14.4	18.1	14.9	15.0	13.3	12.8
	⊕	✓	17.2	12.4	18.3	14.7	18.8	18.6	16.0	13.4	13.3
	✓	✓	23.3	17.2	23.8	19.9	34.3	24.9	30.2	15.2	19.4
wiki	✗	✗	7.8	9.5	8.3	11.4	12.6	9.8	11.2	9.5	9.7
	✗	✓	7.3	8.2	9.8	9.4	10.9	9.7	11.2	9.8	9.3
	⊕	✓	7.3	9.0	14.5	9.8	11.4	9.6	12.2	12.4	9.6
	✓	✓	8.8	10.7	16.9	10.3	12.1	10.9	13.9	13.4	10.1
wiki ∩ bitexte	✗	✗	8.3	9.7	8.4	11.2	12.7	10.0	11.1	9.4	9.5
	✗	✓	8.0	8.4	9.9	9.2	10.5	10.3	11.3	9.8	9.6
	⊕	✓	8.0	8.8	12.6	9.3	11.4	11.9	11.9	10.8	9.7
	✓	✓	12.8	13.2	14.0	12.0	22.4	14.7	20.5	14.7	14.6

TABLE 4: Taux d'erreur (%) obtenus par un modèle CRF partiellement observé sur la tâche d'analyse morpho-syntaxique par transfert cross-lingue, selon que l'on utilise : aucune contrainte (✗) ; les contraintes de type uniquement lorsqu'elles sont différentes des contraintes de référence ⊕ ; ou les contraintes de type pour tous les mots (✓) pour définir l'espace de recherche à l'apprentissage (appr.) et/ou au test (test). Les contraintes de type sont obtenues en combinant un dictionnaire tiré des bitextes (bitexte) et un dictionnaire issue d WIKTIONNAIRE (wiki) (voir le tableau 1 pour le cas de l'union).

l'étiquette NOUN est référence, on peut donc utiliser les contraintes de type et laisser le modèle apprendre à préférer NOUN à VERB uniquement.

Cette stratégie, indiquée par le symbole ⊕ dans le tableau 4, ne permet en fait pas d'améliorer les performances. Au contraire, elle les dégrade pour plusieurs langues. Il semble donc qu'au-delà du fait que les contraintes de type permettent d'accélérer considérablement la vitesse d'apprentissage (d'un facteur 15 environ), elle ne permettent pas de simplifier la tâche du modèle, et même, au contraire, dégradent à nouveau les résultats.

Remarquons enfin que l'impact négatif des contraintes lors de l'apprentissage ne concerne pas seulement les CRF : nous observons le même comportement pour le modèle à base d'historique, HBAL de (Wisniewski *et al.*, 2014), dont la mise à jour est semblable à celle d'un perceptron, entraîné dans les mêmes conditions.

6 État de l'art

L'utilisation de contraintes de type pour l'analyse morpho-syntaxique a surtout été proposé dans le contexte de l'apprentissage non-supervisé (Mérialdo, 1994), que ces contraintes soient extraites de corpus (Goldberg *et al.*, 2008; Ravi & Knight, 2009; Naseem *et al.*, 2009) ou issues de ressources externes comme WIKTIONNAIRE (Li *et al.*, 2012).

Dans le cadre de l'apprentissage supervisé, filtrer les candidats possibles lors du décodage pour accélérer la vitesse de l'analyse morpho-syntaxique est une pratique standard (Ratnaparkhi, 1996; Moore, 2014). Hajic (2000) considère différentes manières d'obtenir des dictionnaires de type pour l'analyse morpho-syntaxique, similaires à nos conditions « corpus », « complétées » et « oracle », et constate que l'utilisation de cette dernière permet d'obtenir davantage de gains que n'en procure un accroissement des données d'apprentissage. Plus récemment, Moore (2014) propose une forme de lissage inspirée du lissage Kneyser-Ney utilisé pour les modèles de langues (Chen & Goodman, 1998), qui permet d'augmenter le rappel des contraintes extraites du corpus, et donc se rapprocher de ce que nous avons appelé les contraintes « corrigées ». À notre connaissance, seuls Smith *et al.* (2005), Waszczuk (2012) et Östling (2013) font état d'utilisation explicite des contraintes lors de l'apprentissage supervisé. Östling (2013) utilise une condition qui serait semblable à ce que nous aurions appelé hapax3.

Smith *et al.* (2005); Waszczuk (2012) séparent l'analyse morpho-syntaxique, pour un jeu d'étiquettes complexe, en deux étapes : une étape de *proposition*, pour laquelle on utilise un module externe proposant un certain nombre d'étiquettes — ce qui revient à construire des contraintes de type ; et une étape de *désambiguïsation*, consistant à prédire en contexte la bonne étiquette parmi les propositions — ce qui revient à effectuer l'apprentissage en incluant les contraintes de type à

l'apprentissage. Pour les tâches considérées, il n'est pas envisageable de se passer des contraintes de type [9].

Müller *et al.* (2013) considèrent une autre manière de filtrer l'espace de recherche, dans le but d'utiliser un modèle CRF d'ordre plus important. Ces auteurs proposent ainsi d'utiliser une cascade de modèles de complexité croissante (Charniak & Johnson, 2005), en réduisant à chaque étape les étiquettes autorisées à chaque position.

Enfin, d'autres manières d'intégrer des contraintes dans un modèle ont été proposées. Dans la régularisation *a posteriori* (Ganchev *et al.*, 2010) la distribution est choisie de manière à maximiser la log-vraisemblance mais également à respecter, en moyenne, certaines contraintes. Un autre cadre permettant d'inclure des contraintes de manière déclarative est celui des modèles conditionnels sous contraintes (Chang *et al.*, 2010, 2012). D'autres manières d'intégrer l'information linguistique dans l'apprentissage supervisé et, plus généralement, de s'interroger sur la meilleure manière de choisir les compétiteurs lors de l'apprentissage, sont l'estimation contrastive (Smith & Eisner, 2005) et ses extensions récentes (Gimpel & Bansal, 2014).

7 Conclusion

Dans cet article, nous avons exploré ce qui nous apparaissait comme un paradoxe, en essayant de répondre à la question suivante : pourquoi l'utilisation de contraintes utiles lors du décodage dégrade les performances lorsque ces mêmes contraintes sont utilisées pour « aider » l'apprentissage ? Nous avons vu que l'intégration des contraintes lors de l'apprentissage conduit à ignorer la contribution de nombreux exemples, à savoir ceux qui seraient pleinement désambiguïsés par les contraintes, et que ceci nuit à la capacité de généraliser à des mots hors-vocabulaire.

Les contraintes de type, permettant d'accélérer considérablement l'apprentissage et le décodage des CRFs, ne peuvent être utilisées telles quelles pendant l'apprentissage sous peine de sévères dégradations des performances, même lorsque ces contraintes sont utilisées lors du décodage. Nous avons proposé quelques pistes permettant de limiter les effets négatifs tout en préservant les bénéfices en temps de calcul, qui est indispensable pour de nombreuses applications.

De manière plus générale, lorsque l'on dispose d'informations linguistiques, il semble important de faire attention à la manière dont on les intègre au modèle, car même une approche en apparence « inoffensive » peut se révéler néfaste pour les performances. En particulier, il peut ne pas être bon d'utiliser cette information, même pertinente lors de l'apprentissage, pour obliger le modèle à tirer au mieux parti du corpus d'entraînement. La meilleure manière d'intégrer une information linguistique externe reste donc une problématique intéressante à étudier.

Références

BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). Tiger : Linguistic interpretation of a german corpus. *Research on Language and Computation*, **2**(4), 597–620.

CHANG M.-W., GOLDWASSER D., ROTH D. & SRIKUMAR V. (2010). Discriminative learning over constrained latent representations. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 429–437, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHANG M.-W., RATINOV L. & ROTH D. (2012). Structured Learning with Constrained Conditional Models. *Machine Learning*, **88**(3), 399–431.

CHARNIAK E. & JOHNSON M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, p. 173–180, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHEN S. F. & GOODMAN J. T. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Rapport interne TR-10-98, Computer Science Group, Harvard University.

FRASER A., SCHMID H., FARKAS R., WANG R. & SCHÜTZE H. (2013). Knowledge sources for constituent parsing of german, a morphologically rich and less-configurational language. *Volume 39, Issue 1 - March 2013*.

GANCHEV K., GRAÇA J. A., GILLENWATER J. & TASKAR B. (2010). Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, **11**, 2001–2049.

9. Même avec celles-ci, Smith *et al.* (2005) indiquent des temps d'apprentissage de plusieurs jours.

GIMPEL K. & BANSAL M. (2014). Weakly-supervised learning with cost-augmented contrastive estimation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1329–1341 : Association for Computational Linguistics.

GOLDBERG Y., ADLER M. & ELHADAD M. (2008). Em can find pretty good hmm pos-taggers (when given a good start). In *Proceedings of ACL-08 : HLT*, p. 746–754 : Association for Computational Linguistics.

HAJIC J. (2000). Morphological tagging : Data vs. dictionaries. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

JURAFSKY D. & MARTIN J. H. (2000). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA : Prentice Hall PTR.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, p. 282–289 : Morgan Kaufmann, San Francisco, CA.

LI S., GRAÇA J. A. V. & TASKAR B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, p. 1389–1398, Stroudsburg, PA, USA.

MÉRIALDO B. (1994). Tagging English text with a probabilistic grammar. *Computational Linguistics*, **20**(2), 155–172.

MOORE R. (2014). Fast high-accuracy part-of-speech tagging by independent classifiers. In *Proceedings of the 25th International Conference on Computational Linguistics : Technical Papers*, COLING'14, p. 1165–1176, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

MÜLLER T., SCHMID H. & SCHÜTZE H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, number October, p. 322–332, Seattle, Washington, USA : Association for Computational Linguistics.

NASEEM T., SNYDER B., EISENSTEIN J. & BARZILAY R. (2009). Multilingual part-of-speech tagging : Two unsupervised approaches. *Journal of Artificial Intelligence Research*, **36**.

ÖSTLING R. (2013). Stagger : an open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, **3**, 1–18.

RATNAPARKHI A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing*, EMNLP'96 : Association for Computational Linguistics.

RAVI S. & KNIGHT K. (2009). Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, p. 504–512, Stroudsburg, PA, USA : Association for Computational Linguistics.

RIEDMILLER M. & BRAUN H. (1993). A direct adaptive method for faster backpropagation learning : The RPROP algorithm. In *Proc. ICNN*, p. 586–591.

SILFVERBERG M., RUOKOLAINEN T., LINDÉN K. & KURIMO M. (2014). Part-of-speech tagging using conditional random fields : Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 259–264 : Association for Computational Linguistics.

SMITH A. N. & EISNER J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 354–362 : Association for Computational Linguistics.

SMITH N. A. (2011). *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

SMITH N. A., SMITH D. A. & TROMBLE R. W. (2005). Context-Based Morphological Disambiguation with Random Fields. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 475–482, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

TÄCKSTRÖM O., DAS D., PETROV S., MCDONALD R. & NIVRE J. (2013). Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, **1**, 1–12.

WASZCZUK J. (2012). Harnessing the CRF Complexity with Domain-Specific Constraints. The Case of Morphosyntactic Tagging of a Highly Inflected Language. In *Proceedings of COLING 2012*, number December 2012, p. 2789–2804, Mumbai, India : The COLING 2012 Organizing Committee.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.