Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue

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

Résumé. Les méthodes de transfert cross-lingue permettent partiellement de pallier l'absence de corpus annotés, en particulier dans le cas de langues peu dotées en ressources linguistiques. Le transfert d'étiquettes morpho-syntaxiques depuis une langue riche en ressources, complété et corrigé par un dictionnaire associant à chaque mot un ensemble d'étiquettes autorisées, ne fournit cependant qu'une information de supervision incomplète. Dans ce travail, nous reformulons ce problème dans le cadre de l'*apprentissage ambigu* et proposons une nouvelle méthode pour apprendre un analyseur de manière faiblement supervisée à partir d'un modèle à base d'historique. L'évaluation de cette approche montre une amélioration sensible des performances par rapport aux méthodes de l'état de l'art pour trois langues sur quatre considérées, avec des gains jusqu'à 3,9% absolus ou 35,8% relatifs.

Abstract. When Part-of-Speech annotated data is scarce, e.g. for under resourced languages, one can turn to cross-lingual transfer and crawled dictionaries to collect partially supervised data. We cast this problem in the framework of *ambiguous learning* and show how to learn an accurate history-based model. This method is evaluated on four languages and yields improvements over state-of-the-art for three of them, with gains up to 3.9% absolute or 35.8% relative.

Mots-clés : apprentissage partiellement supervisé, analyse morpho-syntaxique, transfert cross-lingue.

Keywords: Weakly Supervised Learning, Part-of-Speech Tagging, Cross-Lingual Transfer.

1 Introduction

Les catégories morpho-syntaxiques, qui regroupent les mots partageant un même comportement syntaxique et/ou morphologique, constituent une source d'information pertinente pour de nombreuses tâches de traitement automatique des langues (TAL). Elles sont par exemple aujourd'hui presque systématiquement calculées en prétraitement pour des tâches d'extraction d'information, pour la reconnaissance d'entités nommées ou encore en traduction automatique, sans parler de leur utilisation en analyse syntaxique. Étant donné leur importance, de nombreux travaux se sont attachés à prédire automatiquement ces étiquettes en utilisant une grande variété de méthodes d'apprentissage supervisé. Ces méthodes atteignent aujourd'hui un niveau de performances proche de celui d'un annotateur humain, du moins lorsqu'elles sont entrainées sur des corpus annotés suffisamment grands dans le domaine d'intérêt (Manning, 2011).

L'annotation manuelle d'un corpus reste cependant un processus complexe, fastidieux et onéreux qui nécessite une solide expertise linguistique (Abeillé *et al.*, 2003), même si les outils aujourd'hui disponibles peuvent aider à accélérer très significativement cette démarche (Garrette & Baldridge, 2013). Il n'existe donc actuellement des corpus annotés avec des informations morpho-syntaxiques que pour un nombre de langues et de domaines réduits. Différentes approches ont été proposées dans la littérature pour réduire cet effort d'annotation (voire pour s'en passer complètement) afin de développer des analyseurs morpho-syntaxiques pour des langues et des domaines pour lesquels ces ressources n'existent pas.

Un premier type de solution consiste à estimer des classes de mots automatiquement à partir de corpus non annotés, en

regroupant les unités qui possèdent un même comportement distributionnel ; ces classes doivent ensuite être projetées sur les catégories morpho-syntaxiques traditionnelles pour pouvoir être interprétées. Une grande variété de méthodes ont été proposées dans la littérature pour réaliser cette tâche, depuis (Brown *et al.*, 1992) jusqu'aux travaux plus récents de (Banko & Moore, 2004; Toutanova & Johnson, 2007). Malgré des progrès constants, leurs performances restent en général trop faibles pour permettre leur utilisation dans des applications de TAL (Christodoulopoulos *et al.*, 2010). Cette approche peut être largement améliorée dès lors que l'on dispose d'une poignée de données annotées en plus des données non étiquetées (apprentissage semi-supervisé) : les annotations serviront, par exemple, à initialiser et/ou à désambigüiser les catégories apprises automatiquement.

Il est également possible, pour projeter les mots sur une liste de catégories prédéfinies, d'utiliser des dictionnaires qui contraignent la liste des étiquettes possibles de chaque mot. Ces dictionnaires, qui permettent de réaliser une désambigüisation partielle, s'avèrent très utiles (par exemple dans un cadre de modèle à données latentes), lorsque de grands corpus non annotés sont disponibles pour l'apprentissage (Merialdo, 1994; Banko & Moore, 2004). De tels dictionnaires peuvent aujourd'hui être obtenus automatiquement à relativement bas cout (Li *et al.*, 2012), par exemple à partir des données de projets tels que `Wiktionary` [1].

Le transfert cross-lingue offre une autre manière, complémentaire, de contourner l'absence ou la rareté de données annotées. Le principe du transfert cross-lingues est d'exploiter des corpus de textes parallèles, qui peuvent aujourd'hui être collectés automatiquement en grande quantité (Resnik & Smith, 2003) et d'utiliser ceux-ci pour *transférer* les sorties des outils d'analyse appliqués à une langue *source* riche en données annotées vers une langue *cible* moins bien dotée. Ainsi, en exploitant les alignements automatiques au niveau des mots, il est possible de projeter les étiquettes morpho-syntaxiques des phrases sources vers les phrases cibles (Yarowsky *et al.*, 2001). Dans la lignée de (Das & Petrov, 2011; Li *et al.*, 2012), Täckström *et al.* (2013) a montré qu'il était possible d'apprendre des analyseurs morpho-syntaxiques de bonne qualité de cette manière, si l'information extraite à partir des alignements venait compléter les indications extraites d'un dictionnaire associant à chaque mot un ensemble d'étiquettes morpho-syntaxiques autorisées. Wang & Manning (2014) montrent qu'il est également possible et préférable de transférer les probabilités calculées par les outils d'analyse en langue source plutôt que de projeter uniquement les étiquettes prédites. La méthode de transfert des étiquettes entre langue source et langue cible, ainsi que l'extraction et l'utilisation des dictionnaires, sont détaillées dans la partie 2.

Les deux approches proposées par Täckström *et al.* (2013), permettant d'apprendre à partir de ces deux sources de données (les étiquettes projetées et les dictionnaires), reposent sur des modèles de séquences (HMM et CRF) et sur une généralisation *ad hoc* de leur critère d'apprentissage, afin d'intégrer les différentes sources d'information. Dans ce travail, nous proposons de reformuler le problème du transfert cross-lingue dans le cadre de l'*apprentissage ambigu* (Bordes *et al.*, 2010; Cour *et al.*, 2011) dont l'objectif est d'estimer un classifieur lorsque le système ne peut accéder, lors de la phase d'apprentissage, qu'à un ensemble d'étiquettes possibles dont une seule est juste et non à l'étiquette de référence. À partir des résultats théoriques développés dans (Bordes *et al.*, 2010), nous introduisons une méthode d'apprentissage capable d'apprendre un étiqueteur morpho-syntaxique dans un contexte faiblement supervisé. Ce modèle d'apprentissage est décrit dans la partie 3 et son évaluation sur quatre langues est présentée dans la partie 4.

Le code source et l'ensemble des ressources utilisées dans ce travail sont disponibles à l'url `http://perso.limsi.fr/wisniews/ambiguous`.

2 Création de corpus d'apprentissage par transfert d'étiquettes

L'objectif de ce travail est de développer des étiqueteurs morpho-syntaxiques en s'appuyant sur le transfert d'annotations entre phrases parallèles afin de pouvoir complètement se dispenser, lors de l'apprentissage, de données étiquetées manuellement. Le transfert d'annotations nécessite de définir une correspondance entre étiquettes des langues source et cible, correspondance qui est obtenue dans ce travail en utilisant un ensemble universel d'étiquettes morpho-syntaxiques simples décrit dans la sous-section 2.1. En suivant la méthode proposée par Täckström *et al.* (2013), nous utilisons deux sources complémentaires d'information pour déterminer les étiquettes des différents mots de la langue cible par transfert cross-lingue : un dictionnaire associant à un mot-type donné l'ensemble de ses étiquettes possibles (partie 2.2) et les alignements entre une phrase annotée et sa traduction (partie 2.3). Ces informations sont fusionnées pour étiqueter automatiquement un corpus d'apprentissage (partie 2.4).

1. `http://www.wiktionary.org/`

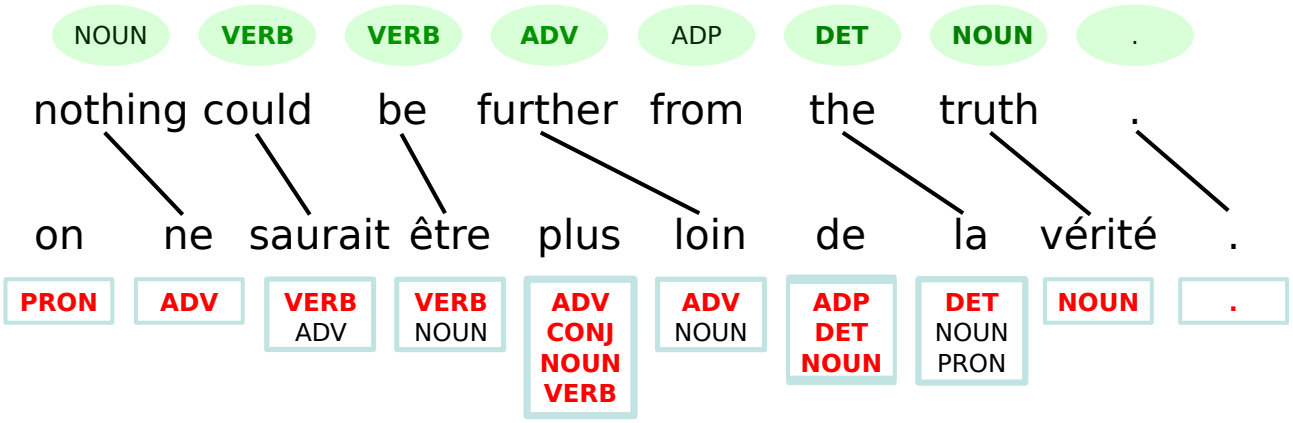

FIGURE 1 – *Exemple de transfert d'étiquettes d'une phrase source (haut) en anglais vers une phrase cible (bas) en français, extrait du corpus d'apprentissage. Pour chaque mot cible, les étiquettes autorisées par les contraintes de type sont représentées dans le cadre bleu. Les étiquettes morpho-syntaxiques de la phrase source sont transférées vers la phrase cible uniquement lorsque celles-ci sont « compatibles » avec les contraintes de type (elles sont dans ce cas représentées en vert). Au final, les étiquettes en rouge constituent l'information (ambigüe) de supervision de la phrase cible.*

2.1 Un ensemble universel d'étiquettes morpho-syntaxique

La possibilité de transférer l'information morpho-syntaxique d'une langue à un autre suppose que cette information puisse être décrite de la même manière dans les deux langues. Même si cette hypothèse forte est hautement controversée (Evans & Levinson, 2009; Broschart, 2009), Petrov *et al.* (2012) définissent 12 étiquettes morpho-syntaxiques à gros grain choisies en raison de leur « universalité » (les catégories identifiées sont relativement stables d'une langue à l'autre) et de leur utilité dans une chaine de traitement de TAL. Ces étiquettes universelles sont les suivantes : NOUN (noms), VERB (verbes), ADJ (adjectifs), ADV (adverbes), PRON (pronoms), DET (déterminants et articles), ADP (prépositions et postpositions), NUM (numéraux), CONJ (conjonctions), PRT (particules), « . » (symboles de ponctuations) et X (pour tout ce qui échappe aux autres catégories, comme par exemple les abréviations ou les mots étrangers). Ces catégories sont uniquement décrites par des exemples et par leur association à des corpus existants et n'ont pas vraiment fait l'objet d'une caractérisation formelle. Par la suite, nous supposerons toujours que toutes les étiquettes morpho-syntaxiques ont été transformées en étiquettes universelles.

2.2 Utilisation de dictionnaires

La première source d'information utilisée pour prédire les informations morpho-syntaxiques est appelée *contrainte de type* et est constituée par un dictionnaire qui associe à chaque mot-type l'ensemble des étiquettes autorisées pour ce mot. La figure 1 donne un exemple d'étiquettes autorisées pour une phrase en français. Les mots « on », « ne », « vérité » et « . » sont entièrement désambigüisés par le dictionnaire. Comme expliqué dans la partie 2.4, ces contraintes permettent de réduire les étiquettes possibles pour chaque mot et de filtrer les annotations transférées en suivant les liens d'alignements.

Plusieurs types de dictionnaires peuvent être utilisés pour déterminer les étiquettes possibles pour un mot. Dans ce travail, nous utilisons des dictionnaires extraits automatiquement de `Wiktionary` en utilisant les méthodes et les heuristiques introduites par Li *et al.* (2012). `Wiktionary` est un dictionnaire de grande envergure, collaboratif et libre et peut être considéré comme une source relativement fiable d'information. Chacune de ses entrées contient les définitions, les étiquettes morpho-syntaxiques et des informations de prononciation, cela pour un grand nombre de mots [2]. Li *et al.* (2012) étudient en détail la couverture et l'exactitude des étiquettes morpho-syntaxiques extraites de `Wiktionary`. Sur les corpus annotés manuellement considérés dans nos expériences, nous observons que ces contraintes de type sont exactes pour plus de 94% des mots-occurrences (voir section 4.1 pour plus de détails). Il est important de noter que les informations de `Wiktionary` sont données en forme libre et que leur extraction et leur conversion vers l'ensemble universel d'étiquettes est une tâche fastidieuse.

2. Par exemple, les dictionnaires français et grec utilisés dans nos expériences contiennent respectivement 1 242 728 et 21 857 entrées. Ces entrées décrivent aussi bien des stemmes que des formes fléchies.

2.3 Transfert cross-lingue d'étiquettes

La deuxième source d'information, appelée *contrainte d'occurrence (token constraint)*, utilise les liens d'alignements [3], lorsqu'ils existent, pour projeter l'étiquette d'un mot-occurrence source sur un mot-occurrence cible. La figure 1 montre un exemple d'alignement entre une phrase source et une phrase cible. L'alignement manquant entre « from » et « de » ne permet pas de transférer l'étiquette ADP. Pour limiter le bruit lié aux erreurs des alignements automatiques, nous calculons au préalable pour chaque mot-type la distribution des étiquettes qui seraient transférées par cette méthode. Pour un mot-occurrence donné, le transfert d'étiquette n'est finalement pris en compte que si cette étiquette transférée est l'une des deux étiquettes les plus fréquentes pour ce mot-type [4]. De plus nous utilisons cette information pour compléter le dictionnaire de types : lorsqu'un mot-type n'est pas dans le dictionnaire extrait de `Wiktionary`, nous utilisons comme contraintes de type ces deux étiquettes les plus fréquentes [5]. Par exemple, sur la figure 1 le verbe conjugué « saurait » n'est pas une entrée de `Wiktionary` et se voit attribuer les deux étiquettes VERB et ADV

Täckström *et al.* (2013) décrivent de manière détaillée l'impact de ces deux types de contraintes et montrent que chacune d'elles apporte des informations complémentaires.

2.4 Prise en compte des deux sources complémentaires d'information

Les deux sources d'information introduites précédemment sont fusionnées en utilisant les règles décrites par l'algorithme 1, qui s'inspire de la méthode de (Täckström *et al.*, 2013). La figure 1 donne un exemple de transfert et de filtrage des étiquettes d'une phrase source vers une phrase cible.

Après transfert et filtrage des étiquettes, les mots cibles sont donc associés à un ensemble d'étiquettes (en rouge, figure 1) et non à une unique étiquette de référence. Un mot-occurrence cible peut cependant être associé à une unique étiquette, dans le cas du transfert d'une étiquette ou dans le cas où la contrainte de type est réduite à une étiquette. La partie 4.1 montre que c'est le cas pour environ 80% des mots-occurrences. Dans la partie suivante, nous expliquons comment il est possible d'entrainer un analyseur morpho-syntaxique n'utilisant que cette *information ambigüe* comme supervision.

Algorithme 1: Règles utilisées pour transférer les étiquettes à partir d'une phrase source.

input : mot w, d dictionnaire décrivant les contraintes de type et un alignement entre les phrases source et cible
output : l'ensemble des étiquettes possibles pour le mot w
$occurrence \leftarrow$ {étiquette du mot avec lequel w est aligné} ; `// ∅ si w n'est pas aligné`
$type \leftarrow d[w]$
if $type \cap occurrence \neq \emptyset$ **then**
  return *occurrence;*
else
  return *type*;
end

3 Modèles de séquences pour l'apprentissage faiblement supervisé

Pour apprendre un modèle de séquences dans un cadre faiblement supervisé, nous utilisons un modèle à base d'historique (Black *et al.*, 1992; Collins, 2003; Tsuruoka *et al.*, 2011) avec une méthode d'apprentissage proche de LaSO (Daumé & Marcu, 2005). Dans les modèles à base d'historique, la prédiction d'une structure complexe (ici la séquence d'étiquettes morpho-syntaxiques) est modélisée sous la forme d'une suite de problèmes de décision, consistant chacun à prédire l'étiquette d'une observation. Chaque décision est prise par un classifieur multi-classe utilisant comme descripteurs des informations extraites de la structure d'entrée, ainsi que les décisions prises antérieurement (c'est-à-dire une sortie partiellement désambigüisée). Ces modèles permettent donc de *réduire* l'apprentissage structuré en un problème d'apprentissage multi-classe.

3. Nous ne considérons que des alignements `1:1` entre mots sources et cibles.
4. Il est aussi possible d'effectuer ce filtrage en seuillant la distribution comme dans (Täckström *et al.*, 2013), mais nous n'avons pas observé de différences entre ces deux heuristiques.
5. Dans les rares cas où un mot-type n'est jamais aligné dans le corpus d'entrainement, nous utilisons le jeu complet d'étiquettes. Dans nos expériences c'est le cas pour moins de 0,2% des mots-occurrences.

Notons $\mathbf{x} = (x_i)_{i=1}^{n}$ la séquence d'observations et $\mathcal{Y}$ l'ensemble des étiquettes possibles (dans notre cas les 12 étiquettes universelles). L'inférence consiste à prédire les étiquettes les unes après les autres en utilisant, ici, un modèle linéaire :

$$y_i^* = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w} | \phi(\mathbf{x}, i, y, h_i) \rangle \tag{1}$$

où $\langle a|b \rangle$ dénote le produit scalaire de a et b, y_i^* est l'étiquette prédite pour la i-ème observation, $\mathbf{w}$ le vecteur de poids, $h_i = y_1^*, ..., y_{i-1}^*$ l'historique décrivant les décisions passées à l'étape i et ϕ un vecteur de traits représentant de manière jointe l'observation, l'étiquette candidate et l'historique. Ainsi, l'inférence peut être vue comme une recherche gloutonne dans l'espace des $\#\{\mathcal{Y}\}^n$ étiquettes possibles pour la séquence observée. Ce type de modèle, qui sacrifie un optimum global au prix d'une plus grande flexibilité des traits [6], a été utilisé avec succès dans de nombreuses applications de TAL (Kazama & Torisawa, 2007; Ratinov & Roth, 2009; Tsuruoka *et al.*, 2011).

L'apprentissage, comme décrit par l'algorithme 2, consiste à effectuer successivement l'inférence pour chaque séquence d'entrée et à corriger le vecteur de poids chaque fois qu'une décision erronée est prise. De manière cruciale (Wolpert, 1992; Ross & Bagnell, 2010), lors de l'apprentissage, les historiques doivent être constitués des étiquettes prédites par le modèle jusque-là, et non des étiquettes de références comme dans (Daumé & Marcu, 2005), afin de rester en cohérence avec la situation qui sera rencontrée au moment du décodage. Cette particularité est la principale différence avec la méthode originale de (Daumé & Marcu, 2005).

L'utilisation d'un modèle à base d'historique permet d'apprendre facilement un modèle de séquences à partir d'une information ambigüe : l'information de supervision disponible est utilisée pour déterminer si une décision est bonne ou erronée ce qui, comme nous allons le montrer dans les deux paragraphes suivants, permet d'adapter la méthode d'apprentissage à un contexte supervisé ou ambigu.

3.1 Apprentissage (fortement) supervisé

Lorsque l'apprentissage est supervisé, la séquence d'étiquettes correcte est connue. Il est donc possible de savoir, à chaque étape de l'inférence, si l'étiquette prédite est différente de l'étiquette de référence. Dès que c'est le cas, la décision est considérée comme erronée et le vecteur de poids est mis à jour, comme pour un perceptron, de la manière suivante :

$$\mathbf{w} \leftarrow \mathbf{w} + \phi\left(\mathbf{x}, i, \hat{y}_i, h_i\right) - \phi\left(\mathbf{x}, i, y_i^*, h_i\right) \tag{2}$$

où y_i^* et $\hat{y}_i$ sont, respectivement, l'étiquette prédite et l'étiquette de référence. Cette mise à jour correspond à un pas de descente de gradient stochastique et permet de renforcer le score de l'étiquette de référence par rapport à tous les autres.

3.2 Apprentissage ambigu (ou faiblement supervisé)

Lorsque la séquence d'étiquettes de référence n'est pas connue, il est tout de même possible d'apprendre un modèle de séquences si l'on dispose pour chaque observation d'un sous-ensemble d'étiquettes possibles, noté $\hat{\mathcal{Y}}_i$. Dans ce cas, une décision est considérée erronée à partir du moment où l'étiquette prédite n'est pas dans l'ensemble des étiquettes autorisées par l'information de supervision ambigüe. Le vecteur de poids est alors mis à jour comme suit :

$$\mathbf{w} \leftarrow \mathbf{w} + \sum_{\hat{y}_i \in \hat{\mathcal{Y}}_i} \left(\phi\left(\mathbf{x}, i, \hat{y}_i, h\right) - \phi\left(\mathbf{x}, i, y_i^*, h_i\right)\right) \tag{3}$$

Cette mise à jour vise à renforcer toutes les étiquettes de $\hat{\mathcal{Y}}_i$.

Dans le cadre de l'apprentissage ambigu (Bordes *et al.*, 2010; Cour *et al.*, 2011), il est possible de montrer en faisant des hypothèses peu restrictives (qui reviennent à dire, en première approximation, qu'il suffit que l'étiquette correcte soit présente dans l'ensemble des étiquettes possibles et qu'elle n'y soit pas systématiquement associée à une autre étiquette), qu'un classifieur entrainé uniquement à partir d'informations de supervision ambigüe revient à un classifieur appris à partir de l'information de supervision complète [7]. Nous nous limiterons ici à donner l'intuition de ce résultat sur un exemple

6. La complexité de l'apprentissage et de l'inférence ne dépend pas de la taille de l'historique, alors que considérer des modèles comme les CRFs avec des dépendances dont l'ordre est supérieur à deux rend aussi bien la complexité de l'apprentissage que l'inférence prohibitive.

7. Bordes *et al.* (2010) définissent une fonction de perte dite *ambigüe (ambiguous loss)*, qui est optimisée par des mises à jour semblables à celles données par l'équation (3) et montrent que la solution qui permet d'obtenir l'erreur minimale pour cette fonction de perte est également la solution du problème minimisant la perte 0/1 que l'on pourrait évaluer si l'on connaissait l'étiquette de référence.

jouet : considérons un corpus contenant deux phrases, « `la souris` » et « `la féline` » dont les étiquettes ambigües sont, respectivement $[\{\text{DET}\}, \{\text{VERB}, \text{NOUN}\}]$ et $[\{\text{DET}\}, \{\text{NOUN}, \text{ADJ}\}]$ et considérons les deux traits correspondant au mot et au mot précédent ; lors de l'application de la règle de mise à jour donnée par l'équation (3), le trait décrivant le mot précédent (« la » dans les deux cas) associé à l'étiquette NOUN sera « renforcé » deux fois, contre une pour les étiquettes incorrectes (ADJ et VERB). Ce « partage de l'information » par l'intermédiaire des traits permet que l'étiquette NOUN soit correctement prédite lors de l'inférence, même si elle n'a jamais été associée aux deux noms de manière non ambigüe. De manière plus générale, tant que deux étiquettes ne sont pas systématiquement associées dans les ensembles de supervision, la répétition des mises à jour renforcera plus souvent la « bonne » étiquette et, au final, celle-ci finira par avoir le plus grand score.

Algorithme 2: Algorithme d'apprentissage. Dans le cas ambigu, $\hat{\mathcal{Y}}_i$ est l'ensemble des étiquettes autorisées ; dans le cas supervisé, cet ensemble est réduit à l'étiquette de référence. Le nombre T d'itérations effectuées, est un hyperparamètre de l'algorithme.

for $t \in [\![1, T]\!]$ **do**
  Tirer au hasard un exemple $\mathbf{x}, \hat{\mathbf{y}}$;
  $h \leftarrow$ liste vide ; `// Initialise un historique vide`
  for $i \in [\![1, n]\!]$ **do**
    $y_i^* = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w} | \phi(\mathbf{x}, i, y, h_i) \rangle$;
    if $y_i^* \notin \hat{\mathcal{Y}}_i$ **then**
      $\mathbf{w} \leftarrow$ mise_à_jour$(\mathbf{w}, \mathbf{x}, i, \hat{\mathcal{Y}}_i, y_i^*, h_i)$; `// Suivant les Équations (2) et (3)`
    end
    ajouter(y_i^*, h);
  end
end

4 Expériences

4.1 Corpus

Nous considérons quatre langues [8] pour évaluer notre approche : le grec, le français, l'espagnol et l'allemand. Dans toutes ces expériences, nous partons de l'anglais comme langue source et utilisons comme données parallèles les corpus EUROPARL et NEWSCOMMENTARY [9]. Pour chaque paire de langues considérée, les corpus sont alignés en utilisant la chaine de traitement standard de MOSES (Koehn *et al.*, 2007) avec l'heuristique d'intersection pour fusionner les deux directions d'alignements. Cette heuristique ne conserve que les liens prédits conjointement dans les deux directions, qui correspondent intuitivement aux alignements les plus surs.

Les étiquettes morpho-syntaxiques pour les phrases sources en anglais sont prédites automatiquement en utilisant un modèle CRF linéaire standard entrainé sur le Penn Treebank. Les étiquettes sont ensuite transférées vers les phrases cibles en utilisant la méthode décrite dans la partie 2. Pour le français et l'espagnol, nous avons réextrait les contraintes de type avec nos propres outils. Pour l'allemand et le grec, nous avons directement utilisé les contraintes extraites par Li *et al.* (2012) [10].

Comme le montrent les statistiques présentées dans le tableau 1, `Wiktionary` fournit des informations pour un grand nombre de mots des corpus d'apprentissage. Cette information reste cependant fortement ambigüe, mais peut être efficacement complétée par les informations extraites des alignements. Pour l'allemand par exemple, considérer conjointement les deux sources d'information permet de réduire le nombre moyen d'étiquettes par mot de 4,6 à 1,1. Ces deux sources d'informations sont complémentaires : le dictionnaire permet de filtrer les étiquettes transférées (entre 10 et 15% des étiquettes ne sont pas transférées en raison de leur incompatibilité avec les contraintes de type) ; en même temps, les étiquettes transférées permettent de lever l'ambigüité des contraintes de type pour environ 50% des mots-occurrences alignés.

8. La quasi-totalité des ressources utilisées dans les travaux antérieurs ne sont pas ou plus distribuées, ce qui complique fortement toute comparaison directe.
9. Sauf pour le grec, pour lequel nous n'utilisons que le corpus EUROPARL.
10. Ces ressources sont disponibles à l'url `https://code.google.com/p/wikily-supervised-pos-tagger/`

Au final l'utilisation conjointe des contraintes et les règles de transfert permettent de créer un corpus d'apprentissage fortement désambigüisé, puisque la plupart des mots-occurrences du corpus d'apprentissage se retrouvent associés à une seule étiquette, et seule une petite partie des mots correspond à plus de 3 étiquettes.

	contraintes	français	grec	espagnol	allemand
% des mots-occurrences dans `Wiktionary`		91,4%	66,0%	87,7%	69,3%
nombre moyen d'étiquettes par mot-occurrence	*wiki*	2,5	5,0	2,8	4,6
nombre moyen d'étiquettes par mot-occurrence	*wiki*$^+$	1,7	1,6	1,6	1,6
nombre moyen d'étiquettes par mot-occurrence	*wiki*$^+$ & *transfert*	1,3	1,1	1,3	1,1
% de mots-occurrences avec une seule étiquette	*wiki*$^+$ & *transfert*	79,4%	88,1%	78,4%	89,5%
% de mots-occurrences avec une ou deux étiquettes	*wiki*$^+$ & *transfert*	90,8%	99,6%	96,6%	99,7%
% de mots-occurrences cibles alignés		71,1%	73,9%	74,0%	69,9%
% d'étiquettes transférées		85,9%	88,2%	85,8%	88,8%
% d'étiquettes transférées informatives		39,0%	50,6%	38,8%	53,7%

TABLE 1 – Statistiques sur l'ambigüité des étiquettes par mot-occurrence dans les corpus parallèles d'apprentissage après filtrage par le dictionnaire extrait de `Wiktionary` (*wiki*), puis lorsque l'on complète celui-ci avec des contraintes de types extraites des alignements (*wiki*$^+$) (section 2.3) et finalement en utilisant la méthode de transfert introduite à la section 2.4 (*transfert*). Le pourcentage d'étiquettes transférées correspond au pourcentage de liens d'alignements pour lesquels l'étiquette transférée est dans les contraintes de type ; le pourcentage d'étiquettes informatives correspond au pourcentage de liens d'alignements pour lesquels l'étiquette transférée est dans les contraintes de type, mais uniquement lorsque ces contraintes de types sont ambigües.

Notre approche est évaluée pour chaque langue considérée sur les ensembles de test des campagnes d'évaluation d'analyse en dépendances (Buchholz & Marsi, 2006; Nivre *et al.*, 2007) [11]. Ces corpus ont été constitués manuellement par des experts linguistes et contiennent plusieurs types d'annotations, dont des étiquettes morpho-syntaxiques fines qui sont transformées en leur équivalent dans le jeu d'étiquettes universelles en utilisant les règles de (Petrov *et al.*, 2012). La qualité des analyseurs entrainés est évaluée par leur taux d'erreur par occurrence sur le jeu de test.

4.2 Traits

Dans toutes nos expériences, nous utilisons des traits similaires à ceux qui sont généralement utilisés dans des tâches d'analyse morpho-syntaxique :

- Pour le mot courant ainsi pour que les deux mots précédents et les deux mots suivants :
  - **identité du mot** : mot en minuscules s'il apparait plus de 10 fois dans le corpus d'apprentissage ;
  - **suffixes** : les suffixes de 2 et 3 lettres s'ils apparaissent dans plus de 20 mot-types différents dans le corpus d'apprentissage ;
  - **classe** : la classe de ce mot [12] parmi 50 classes estimées sur le corpus d'apprentissage en utilisant MKCLS [13]. Les clusters de mots, appris de manière non supervisée, ont déjà étés utilisés comme traits pour améliorer les performances nombreuses de tâches de TAL (Koo *et al.*, 2008; Täckström *et al.*, 2012; Owoputi *et al.*, 2013; Täckström *et al.*, 2013) ;
- **majuscule** : deux traits binaires qui indiquent si le mot courant commence par une majuscule ou non ;
- **trait d'union** : deux traits binaires qui indiquent si le mot courant comporte un trait d'union ou non ;
- **type d'alphabet** : deux traits qui indiquent si le mot est écrit dans un alphabet grec ou latin ;
- **information de structure** : les étiquettes prédites pour les deux mots précédents, la conjonction de ces deux étiquettes, la conjonction de l'étiquette précédente et du mot précédent.

Ces caractéristiques sont semblables à celles sont qui sont utilisées dans (Täckström *et al.*, 2013; Li *et al.*, 2012), exceptées les informations de structure qui ne peuvent être facilement considérées dans un modèle de séquence linéaire comme les CRF.

11. Pour le français, le corpus de test est constitué des 2 000 premières phrases du French Treebank.
12. Les mots hors vocabulaire lors du test sont arbitrairement associés à la classe 1.
13. `http://code.google.com/p/giza-pp/`

4.3 Conditions expérimentales

Pour toutes les paires de langues considérées, un analyseur morpho-syntaxique est entrainé à partir des étiquettes ambigües. Le nombre d'itérations dans l'algorithme 2 est fixé à $T = 100\,000$, ce qui revient à dire que les paramètres de notre méthode sont estimés sur un sous-corpus de 100 000 phrases choisies aléatoirement dans le corpus d'apprentissage. Nos expériences préliminaires indiquent que le choix de ces phrases n'a que peu d'impact sur les performances obtenues. Il apparait également que considérer plus de phrases ne permet pas d'améliorer les performances.

Nous donnons également les résultats pour une réimplémentation du modèle CRF partiellement observé de (Täckström *et al.*, 2013) avec le même jeu de traits que les auteurs, en utilisant 30 itérations de R-Prop sur 100 000 phrases du corpus d'apprentissage et une régularisation ℓ_1 et ℓ_2 [14].

4.4 Résultats

Les résultats obtenus par notre méthode sont résumés dans le tableau 2. Les meilleurs scores de (Täckström *et al.*, 2013) et (Li *et al.*, 2012) pour les langues considérées y sont également inclus, même si ceux-ci ne sont pas directement comparables [15] puisque les différents modèles n'ont pas été entrainés exactement à partir des mêmes ressources (dictionnaires extraits de `Wiktionary`, corpus d'apprentissage, méthode d'alignement, etc).

Ce tableau montre que pour trois des quatre langues, notre méthode améliore sensiblement les résultats de l'état de l'art. L'utilisation de ressources de meilleure qualité, notamment pour les contraintes de type qui ont été extraites à partir d'une version plus récente de `Wiktionary` pour le français et l'espagnol, peut expliquer une partie des gains observés. Cependant, les résultats obtenus sur le grec, pour lequel nous utilisons les mêmes ressources que (Li *et al.*, 2012) et donc plus anciennes que celles utilisées dans Täckström *et al.* (2013), semblent indiquer qu'au moins une partie des améliorations est imputable à la méthode d'apprentissage introduite dans ce travail. De plus en entrainant sur nos ressources un modèle CRF partiellement observé, nous obtenons des résultats comparables à ceux de (Täckström *et al.*, 2013) pour leur modèle équivalent [16], à l'exception de l'allemand. Des expériences supplémentaires sont cependant nécessaires pour déterminer plus précisément les raisons de ces améliorations.

	français	grec	espagnol	allemand
Méthode proposée	**8,9%**	**8,3%**	**7,0%**	10,1%
CRF partiellement observé	13,2%	10,6%	14,0%	18,9%
meilleur score de (Täckström *et al.*, 2013)	11,6%	10,5%	10,9%	**9,5%**
meilleur score de (Li *et al.*, 2012)	—	20,8%	13,6%	14,2%
Inexactitude	5,9%	1,5%	3,4%	3,2%

TABLE 2 – Performances obtenues par notre méthode, un CRF partiellement observé similaire à $\hat{\mathcal{Y}}^{\text{CRF}}_{\text{wik.}}$+C+L dans (Täckström *et al.*, 2013) et les méthodes de l'état de l'art. L'inexactitude est le pourcentage de mots-occurrences dans le corpus de test pour lesquels la contrainte de type n'inclut pas l'étiquette de la référence et correspond à la meilleure performance que pourrait atteindre notre système.

4.5 Discussion

En première analyse, les résultats obtenus par les méthodes de transfert semblent encore très éloignés des performances des meilleurs étiqueteurs morpho-syntaxiques entrainés de manière supervisée. Ainsi, pour l'espagnol, un modèle CRF utilisant les mêmes traits que notre méthode (partie 4.2) appris sur le corpus d'entrainement des données CONLL, atteint un taux d'erreur de seulement 1,3% sur les données de test, contre 7,0% pour notre méthode. Il faut toutefois noter que l'évaluation des approches comme la nôtre comporte un fort biais. En effet, dans la majorité des travaux sur l'étiquetage

14. Dans un contexte de langue cible peu dotée en ressources il n'est pas possible d'utiliser un corpus de développement pour choisir les hyperparamètres des modèles. Comme dans (Täckström *et al.*, 2013), nous fixons arbitrairement les hyperparamètres du CRF partiellement observé à 1.

15. Il faut également noter que Täckström *et al.* (2013) et Li *et al.* (2012) proposent tous deux plusieurs méthodes et que seul le meilleur résultat obtenu sur le corpus de test (et non sur un corpus de validation) est présenté.

16. Sur le tableau 2, seul le meilleur modèle de (Täckström *et al.*, 2013) est indiqué. Le modèle CRF partiellement observé est leur meilleur modèle pour le grec et pour l'allemand.

morpho-syntaxique, l'évaluation est réalisée sur des corpus du même domaine que les corpus d'entrainement, comme pour le modèle CRF introduit ci-dessus. Les méthodes exploitant un transfert bilingue, en revanche, reposent sur des corpus d'apprentissage parallèles, qui peuvent être plus ou moins proches du corpus de test. Par ailleurs, les données de test utilisées exploitent une segmentation en mots qu'il n'est pas toujours aisé de reproduire à l'apprentissage et les conventions d'étiquetage ne sont pas nécessairement les mêmes que celles qui sont utilisées lors de l'apprentissage. Si le premier problème n'a qu'un impact limité sur les performances (il ne concerne que des mots isolés et n'a donc pas d'impact systématique) le second soulève un problème plus fondamental de notre approche ou, du moins, de son évaluation.

L'étiquetage d'un corpus repose sur des conventions qui peuvent varier d'une campagne d'annotation à une autre. Si ces conventions ne sont pas les mêmes pour les corpus de test et d'apprentissage, les prédictions seront entachées d'erreurs systématiques et l'estimation des performances sera biaisée. La situation est encore plus compliquée dans le cadre du transfert d'annotations dans lequel les méthodes utilisent généralement plusieurs sources de données (dans notre cas : `Wiktionary`, le corpus parallèle et le corpus de test) dont les étiquettes doivent toutes être mises en correspondance avec le jeu d'étiquettes universelles. Ce problème est exacerbé dans l'évaluation des méthodes de transfert, dans la mesure où les étiquettes du corpus d'apprentissage sont transférées à partir d'un corpus construit indépendamment de celui utilisé pour l'évaluation. Par exemple, dans le corpus français issu du French Treebank utilisé lors de notre évaluation, les nombres sont étiquetés soit comme des déterminants (DET), par exemple dans le fragment « Christian Blanc, 44 ans » ou « un prêt de 25 millions de dollars », soit comme des adjectifs (ADJ), comme dans « le Monde du 12 janvier » ou « à la page 23 ». Dans le Penn Treebank en revanche, sur lequel sont apprises les étiquettes de la langue source qui seront transférées sur la langue cible, les nombres sont systématiquement associés à l'étiquette NUM. Nous pensons que cette différence est davantage due à un choix de convention qu'à une réalité linguistique. De la même manière, dans le corpus de test pour l'espagnol, *poco* (peu) est majoritairement étiqueté comme un déterminant alors que dans le Penn Treebank, *few* est systématiquement étiqueté comme un adjectif et que `Wiktionary` identifie *poco* soit comme un pronom, un adjectif ou un nom.

Pour évaluer l'impact des différences de conventions d'annotation ainsi que de l'effet du changement de domaine il faudrait pouvoir entrainer un analyseur morpho-syntaxique de manière supervisée sur les données parallèles utilisées lors de l'apprentissage de notre méthode faiblement supervisée. Comme il n'existe pas, à notre connaissance, de données parallèles étiquetées morpho-syntaxiquement, nous avons créé un tel corpus de manière artificielle en étiquetant automatiquement les phrases cibles du corpus parallèle espagnol à l'aide d'un analyseur en catégories morpho-syntaxiques état de l'art [17]. Un CRF entrainé sur ces données obtient un taux d'erreur sur le corpus de test de $6{,}7\%$. Cette valeur, proche de celle obtenue par transfert, montre bien que la principale source d'erreur de notre méthode est liée aux différences de domaine et de convention d'annotation et non à des limites intrinsèques de la méthode d'apprentissage ou de transfert.

Il faut également considérer, dans l'analyse des performances obtenues, que les systèmes faiblement supervisés utilisant les contraintes de type sont fortement limités par l'exactitude de ces contraintes (cf. tableau 2). À titre d'exemple, pour l'espagnol, si l'on contraint le CRF appris de manière supervisée sur le corpus d'apprentissage de CoNLL à ne choisir que des étiquettes autorisées par les contraintes de type, sa performance chute de $1{,}3\%$ à $4{,}3\%$. L'exactitude des contraintes de type dépend elle aussi largement des conventions d'annotation. Pour le français, le taux élevé d'inexactitude s'explique principalement par un faible nombre de mots-types très fréquents (par exemple « au » ou « du ») dont les étiquettes morpho-syntaxiques diffèrent de manière systématique entre `Wiktionary` et le corpus de test.

L'évaluation de notre méthode, et plus généralement des méthodes faiblement supervisées, pose donc de nombreux problèmes méthodologiques, qui sont pourtant rarement discutés. L'interprétation des résultats obtenus doit être faite avec précaution, en particulier lorsqu'il s'agit de les comparer avec des méthodes supervisées.

5 Conclusion

Nous considérons dans ce travail le problème de l'apprentissage d'un analyseur morpho-syntaxique lorsque les étiquettes de supervision ne sont que partiellement connues, par exemple lorsque celles-ci sont automatiquement transférées à partir d'une langue source plus riche en annotations. En abordant ce problème sous l'angle de l'apprentissage ambigu, nous montrons qu'il est possible d'étendre un modèle à base d'historique capable d'apprendre dans un contexte faiblement supervisé. Pour trois des quatre langues considérées, notre méthode améliore sensiblement les résultats les plus récents. Enfin, nous discutons des difficultés et des limites que l'évaluation de telles méthodes pose. En particulier, les différences

17. Nous avons utilisé dans nos expériences FREELING (`http://nlp.lsi.upc.edu/freeling/`)

de convention entre différents corpus annotés peuvent largement biaiser les résultats. Il apparait au final, que la mise en œuvre et l'évaluation des méthodes de transfert nécessitent tout de même un effort conséquent et un minimum de connaissances des langues mises en jeu, ce qui conduit à en relativiser en quelque sorte l'intérêt.

Remerciements

Nous tenons à remercier nos relecteurs anonymes pour leurs très nombreux commentaires ainsi que Thomas Lavergne pour l'implémentation du CRF partiellement supervisé.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*. Dordrecht : Kluwer.

BANKO M. & MOORE R. C. (2004). Part of speech tagging in context. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA : Association for Computational Linguistics.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

BORDES A., USUNIER N. & WESTON J. (2010). Label ranking under ambiguous supervision for learning semantic correspondences. In *ICML*, p. 103–110.

BROSCHART J. (2009). Why Tongan does it differently : Categorial distinctions in a language without nouns and verbs. *Linguistic Typology*, **1**, 123–166.

BROWN P. F., DESOUZA P. V., MERCER R. L., PIETRA V. J. D. & LAI J. C. (1992). Class-based n-gram models of natural language. *Comput. Linguist.*, **18**(4), 467–479.

BUCHHOLZ S. & MARSI E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, p. 149–164, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHRISTODOULOPOULOS C., GOLDWATER S. & STEEDMAN M. (2010). Two decades of unsupervised POS induction : How far have we come ? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 575–584, Stroudsburg, PA, USA : Association for Computational Linguistics.

COLLINS M. (2003). Head-driven statistical models for natural language parsing. *Comput. Linguist.*, **29**(4), 589–637.

COUR T., SAPP B. & TASKAR B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, **12**, 1501–1536.

DAS D. & PETROV S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 600–609, Stroudsburg, PA, USA : Association for Computational Linguistics.

DAUMÉ, III H. & MARCU D. (2005). Learning as search optimization : Approximate large margin methods for structured prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, p. 169–176, New York, NY, USA : ACM.

EVANS N. & LEVINSON S. C. (2009). The myth of language universals : Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, **32**, 429–448.

GARRETTE D. & BALDRIDGE J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 138–147, Atlanta, Georgia : Association for Computational Linguistics.

KAZAMA J. & TORISAWA K. (2007). A new perceptron algorithm for sequence labeling with non-local features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 315–324.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, p. 177–180, Stroudsburg, PA, USA : Association for Computational Linguistics.

KOO T., CARRERAS X. & COLLINS M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08 : HLT*, p. 595–603, Columbus, Ohio : Association for Computational Linguistics.

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 : Association for Computational Linguistics.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In *Proceedings of the Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, p. 171–189 : Springer.

MERIALDO B. (1994). Tagging english text with a probabilistic model. *Comput. Linguist.*, **20**(2), 155–171.

NIVRE J., HALL J., KÜBLER S., MCDONALD R., NILSSON J., RIEDEL S. & YURET D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, p. 915–932, Prague, Czech Republic : Association for Computational Linguistics.

OWOPUTI O., O'CONNOR B., DYER C., GIMPEL K., SCHNEIDER N. & SMITH N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 380–390, Atlanta, Georgia : Association for Computational Linguistics.

PETROV S., DAS D. & MCDONALD R. (2012). A universal part-of-speech tagset. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

RATINOV L. & ROTH D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, p. 147–155, Stroudsburg, PA, USA : Association for Computational Linguistics.

RESNIK P. & SMITH N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, **29**(3), 349–380.

ROSS S. & BAGNELL D. (2010). Efficient reductions for imitation learning. In *AISTATS*, p. 661–668.

TÄCKSTRÖM O., MCDONALD R. & USZKOREIT J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 477–487, Stroudsburg, PA, USA : Association for Computational Linguistics.

TOUTANOVA K. & JOHNSON M. (2007). A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *NIPS*.

TSURUOKA Y., MIYAO Y. & KAZAMA J. (2011). Learning with lookahead : Can history-based models rival globally optimized models ? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, p. 238–246, Portland, Oregon, USA : 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.

WANG M. & MANNING C. D. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transaction of the ACL*, **2**(1), 55–66.

WOLPERT D. H. (1992). Stacked generalization. *Neural Networks*, **5**.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.