Utilisation des fonctions de croyance pour l'estimation de paramètres en traduction automatique

Christophe Servan Simon Petitrenaud
LIUM, Le Mans
christophe.servan@lium.univ-lemans.fr
simon.petit-renaud@lium.univ-lemans.fr

RÉSUMÉ

Cet article concerne des travaux effectués dans le cadre du 7ème atelier de traduction automatique statistique et du projet ANR COSMAT [1]. Ces travaux se focalisent sur l'estimation de paramètres contenus dans une table de traduction. L'approche classique consiste à estimer ces paramètres à partir de fréquences relatives d'éléments de traduction. Dans notre approche, nous proposons d'utiliser le concept de masses de croyance afin d'estimer ces paramètres. La théorie des fonctions de croyances est une théorie très adaptée à la gestion des incertitudes dans de nombreux domaines. Les expériences basées sur notre approche s'appliquent sur la traduction de la paire de langue français-anglais dans les deux sens de traduction.

ABSTRACT

Feature calculation for Statistical Machine Translation by using belief functions

In this paper, we consider the translation of texts within the framework of the 7th Workshop of Machine Translation evaluation task and the COSMAT corpus using a statistical machine translation approach. This work is focused on the translation features calculation of the phrase contained in a phrase table. The classical way to estimate these features are based on the direct computation counts or frequencies. In our approach, we propose to use the concept of belief masses to estimate the phrase probabilities. The Belief Function theory has proven to be suitable and adapted for the management of uncertainties in many domains. The experiments based on our approach are focused on the language pair English-French.

MOTS-CLÉS : Traduction automatique statistique, fonctions de croyance, apprentissage automatique, estimation de paramètres.

KEYWORDS: Statistical machine Translation, belief function, machine learning, feature estimation.

1 Introduction

Il est classique d'utiliser une table de traduction comme élément d'un modèle de traduction automatique statistique (TAS). Dans un système de traduction automatique fondé sur des segments (ou séquences de mots), une table de traduction contient les traductions alternatives et ses probabilités pour un segment en une langue source donnée. Chaque ligne ou événement d'une table de traduction comprend deux segments, l'un en langue source et l'autre en langue cible.

1. http://www.cosmat.fr

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 543–550,
Grenoble, 4 au 8 juin 2012.

On suppose que les événements sont indépendants les uns des autres. Les tables de traduction peuvent contenir beaucoup de paramètres comme les probabilités de traduction des segments ou les probabilités lexicales. Afin d'estimer ces paramètres, les systèmes de TAS utilisent de grands corpus appelés bitextes, qui sont composés de phrases en langue source et en langue cible qui sont la traduction l'une de l'autre. Pour chaque phrase, les mots des deux langues sont alignés en fonction de la traduction.

Dans l'approche classique, l'estimation des probabilités est faite par l'utilisation des fonctions de compte simples, sur la base de fréquences relatives. Dans de nombreux travaux, la théorie des fonctions de croyance (initialement théorie de Dempster-Shafer) permet une représentation à la fois plus souple et plus précise de différents types d'incertitude que les modèles probabilistes (Smets, 1988; Cobb et Shenoy, 2006). Par exemple, les modèles probabilistes peuvent difficilement prendre en compte le conflit entre deux hypothèses différentes de traduction, en particulier dans le cas des exemples rares. Il est également délicat d'estimer le degré de confiance global que l'on a sur l'ensemble des éléments de traduction pour une source donnée.

La théorie des fonctions de croyance est capable de traiter ce genre de situations en fournissant des degrés de conflit quand il y a des hypothèses contradictoires, ainsi que des mesures de confiance globale. Dans cet article, nous proposons une méthode originale pour estimer les paramètres associés aux événements constitués de paires de segments à l'aide des fonctions de croyance.

Cet article présente nos premiers travaux et leurs résultats réalisés avec cette nouvelle approche. Il est organisé comme suit : tout d'abord, nous rappelons brièvement la théorie de la traduction automatique statistique. Dans la section 3, nous détaillons notre approche basée sur les fonctions de croyance. Ensuite, nous présentons des expériences sur différents corpus de traduction français-anglais. Enfin, nous concluons cet article et proposons quelques perspectives.

2 Estimation de paramètres en traduction automatique statistique

Soit une phrase source s traduite en un certain nombre de phrases cibles $t \in T_s$, où T_s est l'ensemble de toutes les traductions observées de s dans la table de traduction. Le modèle de traduction automatique statistique (TAS) utilise un ensemble de n fonctions $f_i, i = 1 \ldots, n$, qui dépendent des séquences de mots sources et cibles, afin de déterminer la meilleure traduction. Les fonctions que l'on considère habituellement comprennent le modèle de traduction, le modèle de distorsion, le modèle de langage cible et différentes pénalités. Parmi toutes les traductions possibles, celle qui sera choisie maximise la probabilité *a posteriori*, et peut s'exprimer de la façon suivante :

$$t^* = \arg \max_{t=1 \in T_s} \log \left(\prod_{i=1}^{n} f_i(t,s)^{\lambda_i} \right), \quad (1)$$

où chaque paramètre λ_i est un coefficient de pondération pour chaque fonction f_i (Och, 2003). Ces poids sont généralement optimisés de façon à maximiser la performance de traduction sur

langue source (s) - fr	langue cible (t) - en
...	...
étant donné un	given a
étant donné un	starting from an
étant donné	given
étant donné	given
étant donné	starting from
étant donné	starting
étant	starting
...	...

Tableau 1 – Exemple de paires de segments extraits d'un bitexte

des ensembles de données de développement. Le travail présenté dans cet article se focalise sur les caractéristiques utilisées pour estimer le modèle de traduction.

Dans l'outil de traduction classique « Moses » (Koehn *et al.*, 2007), la table de traduction contient cinq caractéristiques (Koehn, 2010) : les paramètres de traduction des segments, la pondération lexicale des traductions et la pénalité des segments. Les paramètres de traduction des segments sont estimés à l'aide de la règle de décision de Bayes dans les deux sens de traduction. Les poids lexicaux sont estimés à partir du modèle IBM 1 basés sur les mots de chaque paire de segment. Enfin, la pénalité d'apparition du segment est définie par l'utilisateur. Cette fonction permet de privilégier les segments en fonction de leur longueur et prend une valeur constante ρ pour tous les segments. Si $\rho > e$, on préférera des segments longs aux segments courts. Inversement, si $\rho < e$, les segments courts seront privilégiés. Une fois que ces paramètres sont définis, les poids de l'ensemble de ces caractéristiques sont optimisés au cours du processus d'entraînement par le taux d'erreur minimum (MERT) (Och, 2003).

Dans cet article, nous nous concentrons sur l'estimation des paramètres associés aux segments de traduction dans les deux sens de traduction, nous estimons les probabilités lexicales de manière classique et, enfin, nous fixons la pénalité d'apparition ρ à la valeur e.

Le tableau 1 donne un exemple de paires de segments extraits d'un bitexte. A partir de cet exemple, nous obtenons la table de traduction présenté dans le tableau 2 contenant l'estimation des différentes probabilités. Ainsi, la probabilité de traduction de « starting » sachant « étant donné » est une simple fréquence conditionnelle égale à 0,25 et la probabilité de « given » sachant « étant donné » est égale à 0,5. La probabilité conditionnelle inverse du segment de traduction est estimée de la même manière.

Cette façon d'estimer les paramètres a quelques inconvénients. Lorsque certaines paires de segments apparaissent plusieurs fois, comme la paire « *la maison blanche*|*the white house* », et n'ont pas d'occurrences concurrentes, l'estimation de la probabilité du segment est égale à 1, mais dans d'autres situations, des événements peuvent survenir très rarement et être ambigus. Par exemple, supposons que pour le mot français « *chien* » (qui devrait se traduire par « *dog* » en anglais),

segment source (s) - fr	segment cible (t) - en	$p(t\|s)$	$lex(t\|s)$	$p(s\|t)$	$lex(s\|t)$	ρ
...	...					
étant donné	given	0,5	0,060147	0,333333	0,306373	2,718
étant donné	starting	0,25	7,15882e-06	0,333333	5,19278e-05	2,718
étant donné	starting from	0,25	7,15882e-06	0,333333	0,0277778	2,718
...	...					

Tableau 2 – Exemple de table de traduction avec les différents paramètres

deux occurrences contradictoires soient disponibles dans la table de traduction : « $chien|cat$ » et « $chien|dog$ ». Pour chacun de ces deux événements, l'estimation de la probabilité peut être égale à 1, car ils n'ont été observés qu'une seule fois. Par exemple, dans le cadre de notre expérience sur le corpus COSMAT, il existe 13 480 cas correspondant à 33 900 entrées sur les 363 324 que compte la table de traduction, soit un peu moins de 10%, dans le sens de traduction français-anglais.

Même si l'estimation de la probabilité de la traduction des paires *inversées* « $cat|chien$ » et « $dog|chien$ » peut équilibrer ce problème, si l'événement n'est observé qu'une seule fois dans les deux sens de traduction, l'estimation des probabilités conditionnelles inversées est inutile. Il existe la possibilité de lisser les probabilités de l'ensemble des événements (Foster *et al.*, 2006). Cependant, les approches de lissage optent pour une redistribution des estimations afin de donner, notamment, une probabilité non nulle aux événements non observés (Chen et Goodman, 1996; Goodman, 2001). Notre but n'est pas celui-ci, mais plutôt de proposer une approche différente de l'estimation des paramètres des événements observés.

L'utilisation de théories alternatives à la théorie des probabilités permet de mieux ajuster ces estimations. L'une d'elles est particulièrement adaptée à la gestion de différents types d'incertitudes : la théorie des fonctions de croyance, qui a été proposée puis développée depuis une trentaine d'années. Cette théorie a été appliquée avec succès à de nombreux domaines tels que l'identification du locuteur (Petitrenaud *et al.*, 2010) ou la classification en général (Elouedi *et al.*, 2000). Dans nos travaux, nous nous utilisons certains concepts fondamentaux de cette théorie pour notre problème d'estimation de paramètres.

3 Fonctions de croyances pour les systèmes de TAS

Dans cette section, nous présentons brièvement quelques notions de la théorie des fonctions de croyance (Shafer, 1976; Smets et Kennes, 1994) et nous l'appliquons au problème d'estimation de paramètres de modèles de traduction. Dans cet article, nous adoptons le point de vue proposé par Smets : le modèle de croyances transférables (MCT) (Smets et Kennes, 1994). L'objectif de ce modèle est de déterminer la croyance concernant différentes propositions, à partir d'informations disponibles.

Soit Ω un ensemble fini, appelé cadre de discernement de l'expérience. La représentation de l'incertitude est faite par le biais de la notion de fonction de croyance, définie comme une fonction m de 2^Ω sur $[0,1]$ telle que $\sum_{A\subseteq\Omega} m(A) = 1$. La quantité $m(A)$ représente la croyance allouée à la proposition A, et à aucune proposition plus restrictive. Une des opérations les plus importantes dans le MCT est la procédure d'agrégation des informations, c'est-à-dire la combinaison de plusieurs fonctions de croyance définies dans un même cadre de discernement (Smets et Kennes, 1994). En particulier, la combinaison de deux fonctions de croyance m_1 et m_2 indépendantes définies sur Ω est faite en utilisant l'opérateur binaire conjonctif $\cap$, tel que $m' = m_1 \cap m_2$ (Smets et Kennes, 1994) :

$$\forall A \subseteq \Omega, m'(A) = \sum_{B\cap C=A} m_1(B)m_2(C) \tag{2}$$

Cet opérateur est associatif et commutatif, il est alors possible de définir la combinaison de n fonctions $m_1, \ldots, m_n$ sur Ω par la fonction de croyance $m = m_1 \cap \ldots \cap m_n$. Cette dernière fonction m capture l'information globale sur l'ensemble des expériences connues.

Ici, nous proposons d'utiliser le MCT pour estimer les paramètres de traduction des segments. Tout d'abord, pour une source s, chaque cible $t_i \in T_s$ donne une information particulière pour la traduction qui peut être décrite par une fonction de croyance m_s^i, telle que :

$$\begin{cases} m_s^i(\{t_i\}) = p(t_i|s) \\ m_s^i(T_s) = \overline{p(t_i|s)} \end{cases}, \tag{3}$$

où $\overline{p(t_i|s)} = 1 - p(t_i|s)$. Si nous combinons les informations définies par toutes les hypothèses disponibles dans la table concernant la traduction de s, à partir de l'opérateur conjonctif défini dans l'équation 2, nous obtenons alors une fonction de croyance $m_s = \cap_{t \in T_s} m_s^i$. La masse de t_i est obtenue par la formule suivante :

$$m_s(\{t_i\}) = p(t_i|s). \prod_{t_k \in T_s \setminus \{t_i\}} \overline{p(t_k|s)}. \tag{4}$$

Notons que généralement $\sum_{t_i \in T_s} m_s(\{t_i\}) = 1 - m(T_s) - m(\emptyset) < 1$. Les masses $m(T_s)$ et $m(\emptyset)$ peuvent être respectivement interprétées comme le degré d'ignorance et le degré de conflit d'informations concernant la traduction de s. Même si $m(\emptyset)$ n'entre pas directement dans notre modèle de traduction, quand il y a un conflit important entre plusieurs hypothèses de traduction, les masses de croyance sur chacun des singletons $t_i \in T_s$ s'affaiblissent. Nous obtenons alors une estimation de la fonction définie dans l'équation 2 par : $f(t_i, s_j) = m_{s_j}(\{t_i\})$. De la même manière, l'estimation de la fonction inverse est obtenue par l'équation suivante :

$$m_t^I(\{s_j\}) = p(s_j|t). \prod_{s_k \in S_t \setminus \{s_j\}} \overline{p(s_k|t)}\,, \tag{5}$$

où S_t est l'ensemble des sources possibles de la cible t. Si nous appliquons ces formules à l'exemple du tableau 1, une nouvelle estimation des paramètres associés aux différentes paires de segments est calculée dans le tableau 3.

$m_s(\text{starting}) = p(\text{starting}\|\text{étant donné}) \cdot \overline{p(\text{given}\|\text{étant donné})} \cdot \overline{p(\text{starting from}\|\text{étant donné})}$
$m_s(\text{starting}) = 0,09375$
$m_s(\text{starting from}) = 0,09375$
$m_s(\text{given}) = 0,28125$
$m_s(T_s) = 0,28125$
$m_s(\emptyset) = 0,25$

Tableau 3 – Exemple d'estimation de paramètres de paires de segments à l'aide du MCT ($s =$ « étant donné »)

Notons que si $p(t_i|s) = 1$, les masses de croyance pour les autres hypothèses deviennent nulles (cf. équation 4). La masse de croyance indiquée dans cette équation peut alors être modifiée de

corpus	AbsTrain		AbsDev		AbsTest		nc7		eparl7		nwtst2010		nwtst2011	
langue	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en
# de phrases	5141		1083		1102		137k		2M		2489		3003	
# de mots	135K	120K	28K	25K	28K	25K	4M	3,4M	61,7M	55,7M	62k	70k	75k	84,5k

Tableau 4 – Description des bitextes.

façon suivante :

$$m_s(\{t_i\}) = \frac{1}{1+\frac{1}{|s|}}\ , \tag{6}$$

où $|s|$ désigne le nombre d'occurrences de s. Ainsi, $m_s(\{t_i\}) < 1$ mais plus on a d'information sur s, plus $m_s(\{t_i\})$ tendra vers 1. Enfin, les phrases cibles choisies sont obtenues par le processus de décision défini par l'équation 2.

4 Expériences

		nc7		eparl7-nc7	
	approche	BLEU	TER	BLEU	TER
Sens de la traduction : fr↦en					
newstest2010	prob.	**24,58 (0,13)**	**57,53 (0,03)**	**27,22 (0,05)**	**57,52 (0,10)**
	MCT	24,56 (0,08)	57,66 (0,07)	27,10 (0,10)	57,74 (0,10)
newstest2011	prob.	**25,92 (0,11)**	**54,48 (0,09)**	**29,52 (0,12)**	**55,08 (0,12)**
	MCT	25,83 (0,17)	54,61 (0,08)	29,47 (0,14)	55,28 (0,13)
Sens de la traduction : en↦fr					
newstest2010	prob.	24,75 (0,06)	60,17 (0,26)	**28,04 (0,07)**	**53,77 (0,14)**
	MCT	**24,74 (0,04)**	**60,07 (0,18)**	28,00 (0,03)	53,76 (0,03)
newstest2011	prob.	26,84 (0,19)	57,75 (0,29)	28,60 (0,25)	52,85 (0,34)
	MCT	**26,93 (0,09)**	**57,63 (0,17)**	**28,60 (0,04)**	**52,74 (0,04)**

Tableau 5 – Résultats obtenus suivant les métriques BLEU et TER avec deux systèmes entraînés sur les corpus : News-Commentary 7 (*nc7*) ; Europarl 7 - News-Commentary 7 (*eparl7-nc7*).

Afin de valider notre méthode, plusieurs expériences ont été réalisées. Tout d'abord, nous avons utilisé le corpus COSMAT, qui est un ensemble de bitextes de résumés de thèses de doctorat en français et en anglais. Puis, nos expériences ont été placées dans le contexte de l'évaluation du septième atelier sur la traduction automatique statistique (WMT12).

4.1 Le Corpus COSMAT

Le projet ANR COSMAT est composé de nombreux résumés de thèse de doctorat en français et en anglais. Ces résumés ont été classés en fonction de plusieurs thèmes. Dans nos expériences, nous n'avons retenu que le domaine associé à l'informatique. Les corpus d'apprentissage, de développement et de tests sont décrits dans le tableau 4.

Sur le corpus de développement, la perplexité des modèles de langage cible est de 122 pour le français et de 196 pour l'anglais. Les modèles sont adaptés à la tâche grâce à l'utilisation du corpus d'entraînement (AbsTrain) et des modèles de langage.

Sens de traduction		fr→en		en→fr	
corpus	approche	BLEU	TER	BLEU	TER
AbsDev	prob.	34,78 (0,09)	48,24 (0,29)	**32,28 (0,02)**	52,82 (0,25)
	belief	**34,85 (0,06)**	**48,25 (0,11)**	32,28 (0,01)	**52,40 (0,18)**
AbsTest	prob.	40,03 (0,44)	44,35 (0,12)	**38,80 (0,19)**	47,76 (0,36)
	belief	**40,44 (0,10)**	**44,18 (0,12)**	38,43 (0,12)	**47,66 (0,10)**

Tableau 6 – Résultats obtenus avec le corpus COSMAT suivant les métriques BLEU et TER.

4.2 Le corpus WMT12

Le cadre utilisé pour l'évaluation de WMT12 contient plusieurs corpus. Ceux que nous avons utilisés dans nos expériences sont décrits dans le tableau 4. Les corpus d'apprentissage sont Europarl 7 (eparl7) et News-Commentary 7 (nc7). Les modèles employés quand la langue cible est le français et l'anglais ont respectivement une perplexité de 123 et de 169.

4.3 Résultats

Les tableaux 6 et 5 contiennent les résultats obtenus avec l'approche classique et avec notre approche basée sur les fonctions de croyance. Les métriques utilisées sont le score BLEU (Papineni *et al.*, 2002) et la métrique TER (Snover *et al.*, 2005). Afin de garantir une certaine robustesse des résultats, trois optimisations de MERT ont été faites. Le résultat présenté correspond à une moyenne de ces trois optimisations et la valeur indiquée entre parenthèses est l'écart-type. La pénalité de brièveté (ou de longueur de phrase) associée au score BLEU est d'environ 0,99 (0,01) pour les deux approches, dans les deux sens de traductions et pour chacune des expériences.

Les expériences menées sur COSMAT et sur WMT12 montrent que notre nouvelle approche semble avoir des résultats similaires à ceux de l'approche classique. Toutefois, le score BLEU a tendance à être plus faible dans notre approche lorsque le sens de la traduction est de l'anglais vers le français dans l'expérience avec le corpus WMT12 mais à l'inverse, dans l'expérience COSMAT, notre nouvelle approche est légèrement moins performante dans le sens français vers anglais. Malgré ce constat, ces premiers résultats sont encourageants et nous poussent à poursuivre dans cette direction.

5 Conclusions et perspectives

Cet article présente les premiers résultats sur l'utilisation du Modèle des Croyances Transférables (MCT) en traduction automatique statistique. Cette théorie a été utilisée pour estimer différemment les paramètres des paires de segments de traduction. Les résultats obtenus dans la traduction français-anglais, dans les deux directions, sur les corpus COSMAT et WMT12 sont encourageants. Prochainement, nous prévoyons d'appliquer le MCT en traduction de manière plus approfondie. D'abord, nous allons étendre cette approche à l'estimation des paramètres de pondération lexicale. Nous allons également orienter nos recherches vers une stratégie de prise en compte de la proximité linguistique des différentes hypothèses de traduction pour une phrase donnée. Pour reprendre l'exemple du tableau 1, « $starting$ » serait notamment plus proche de « $starting\ from$ » que de « $given$ ». Le MCT permet d'intégrer ce genre de situations avec une certaine souplesse.

6 Remerciements

Ce travail a été financé par l'Agence Nationale de la Recherche dans le cadre du projet COSMAT et par la Commission Européenne à travers le projet EUROMATRIXPLUS.

Références

CHEN, S. F. et GOODMAN, J. (1996). An empirical study of smoothing techniques for language modeling. *In* JOSHI, A. et PALMER, M., éditeurs : *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, Californie, États-Unis d'Amérique. Morgan Kaufmann Publishers.

COBB, B. R. et SHENOY, P. P. (2006). A comparison of methods for transforming belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):255–266.

ELOUEDI, Z., MELLOULI, K. et SMETS, P. (2000). Classification with belief decision trees. *In Proceedings of the 9th International Conference on Artificial Intelligence : Methodology, Systems, Architectures.* AIMSA 2000, Springer Lecture Notes on Articial Intelligence.

FOSTER, G., KUHN, R. et JOHNSON, H. (2006). Phrasetable smoothing for statistical machine translation. *In Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 53–61.

GOODMAN, J. T. (2001). A bit of progress in language modeling. *Computer Speech &amp ; Language*, 15(4):403 – 434.

KOEHN, P. (2010). *Statistical MAchine Translation*. Cambridge University Press.

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. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (Demo and Poster Sessions)*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

OCH, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

PETITRENAUD, S., JOUSSE, V., MEIGNIER, S. et ESTÉVE, Y. (2010). Automatic named identification of speakers using belief functions. *In Information Processing and Management of Uncertainty (IPMU'10)*.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

SMETS, P. (1988). Belief functions versus probability functions. pages 17–24.

SMETS, P. et KENNES, R. (1994). The transferable belief model. *Artificial Intelligence*, 66:191–234.

SNOVER, M., DORR, B., SCHWARTZ, R., MAKHOUL, J., MICCIULA, L. et WEISCHEDEL, R. (2005). A study of translation error rate with targeted human annotation. Rapport technique LAMP-TR-126,CS-TR-4755,UMIACS-TR-2005-58, University of Maryland, College Park and BBN Technologies.