Une Étude Empirique de la Capacité de Généralisation des Plongements de Mots Contextuels en Extraction d'Entités

Bruno Taillé[1,2], Vincent Guigue[2], et Patrick Gallinari[2]

[1]BNP Paribas, CIB, Analytics Consulting
[2]Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6

27 mai 2019

Résumé

Les plongements de mots contextuels utilisent la capacité des modèles de langue à tirer parti de données non annotées pour apprendre des représentations de mots dépendant de leur contexte. Ils sont utiles à la généralisation, particulièrement en Reconnaissance d'Entités Nommées où détecter des mentions d'entités jamais rencontrées pendant l'entraînement est crucial. Malheureusement, les benchmarks actuels surestiment l'importance des caractéristiques lexicales par rapport aux caractéristiques contextuelles à cause d'un recoupement lexical non raisonnable entre mentions d'entraînement et d'évaluation. Dans cet article, nous proposons de mieux évaluer la capacité de généralisation des modèles en séparant les mentions par degré de nouveauté ainsi qu'avec une évaluation extra domaine. Nous montrons que les plongements contextuels sont surtout bénéfiques pour la détection des mentions non rencontrées pendant l'entraînement et mènent à une amélioration relative de +9% du score micro-F1 extra domaine contre +2% intra domaine.

Mots-clef : Reconnaissance d'Entités Nommées, Plongements Contextuels, Adaptation de Domaine.

1 Introduction

La Reconnaissance d'Entités Nommées (REN) consiste à détecter les mentions textuelles d'entités et à les classifier selon des types prédéfinies. Cette tâche est modélisée comme de l'étiquetage de séquence dont l'architecture neuronale classique est le BiLSTM-CRF [HXY15]. Les progrès récents de l'état de l'art proviennent principalement de l'utilisation de nouvelles formes de représentations : des plongements de mots appris à l'échelle des caractères [LBS+16] ou des plongements de mots contextuels obtenus par des modèles de langues à l'échelle des mots [PNI+18, DCLT18] ou des caractères [ABV18].

Cette dernière approche d'apprentissage par transfert utilise des représentations de mots apprises par des modèles de langue pour réduire la dépendance aux données annotées spécifiques à une tâche ou un domaine [HR18, RSN+18]. BERT [DCLT18] établit même l'état de l'art actuel avec une simple projection linéaire des états cachés appris par un modèle de langue affiné pour la tâche. Parallèlement, l'adaptation de domaine en REN est souvent limitée au préentraînement d'un modèle sur des données sources dont les prédictions sont utilisées comme entrées d'un second réseau ou bien qui est réentrainé sur les données cibles [LDS18, RCL18].

Dans cet article, nous montrons que les jeux de données CoNLL03 et OntoNotes découragent la généralisation aux entités non rencontrées à cause d'un recoupement lexical non réaliste entre mentions d'entraînement et d'évaluation. Cela conduit à surestimer l'importance des caractéristiques lexicales par rapport aux caractéristiques contextuelles. Aussi, nous proposons de mieux évaluer les capacités de généralisation d'une part en séparant les mentions par degré de nouveauté et d'autre part avec une évaluation extra domaine. Dans ce cadre, nous montrons que les plongements de mots contextuels sont surtout bénéfiques pour la détection des mentions non rencontrées pendant l'entraînement et mènent à une amélioration maximale relative de +9% du score micro-F1 extra domaine contre +2% intra domaine sur CoNLL03. Cela permet d'établir une borne inférieure simple mais efficace d'adaptation de domaine sans données cibles qui pourrait être améliorée en incorporant ces dernières.

	CoNLL03					ON	ON réaligné					ON réaligné / CoNLL03				
	LOC	MISC	ORG	PER	Tous	Tous	LOC	MISC	ORG	PER	Tous	LOC	MISC	ORG	PER	Tous
Exact	82%	67%	54%	14%	52%	67%	87%	93%	54%	49%	69%	70%	78%	18%	16%	42%
Partiel	4%	11%	17%	43%	20%	24%	6%	2%	32%	36%	20%	7%	10%	45%	46%	28%
Nul	14%	22%	29%	43%	28%	9%	7%	5%	14%	15%	11%	23%	12%	38%	38%	30%

TABLE 1 – Recoupement lexical des occurrences de mentions des jeux de test avec les jeux d'entraînement respectifs pour CoNLL03 et OntoNotes original et réaligné. La dernière colonne montre le recoupement entre le test de OntoNotes réaligné et l'entraînement de CoNLL03 dans l'évaluation extra domaine.

2 Jeux de données

CoNLL03 La partie anglaise de CoNLL03 [SDM03] est le benchmark standard en REN et est composé d'articles Reuters datés de 1996 et annotés pour quatre types : Organisation (ORG), Personne (PER), Localité (LOC) et Divers (MISC).

OntoNotes 5.0 OntoNotes 5.0 [WPM$^+$13] est composé de documents de six domaines annotés pour la REN et la Résolution de Coréférence. Il est annoté manuellement pour onze types d'entités et sept types de valeurs qui sont généralement traités sans distinction. La partition entraînement/test classique pour la REN [SVBM17] est la même que celle de la tâche de Résolution de Coréférence de CoNLL-2012 [PMX$^+$12].

3 Recoupement Lexical

Les modèles de REN neuronaux reposent sur des caractéristiques lexicales sous la forme de plongements de mots, qu'ils soient appris au niveau des caractères ou non. Bien que la syntaxe soit incorporée par leur préentraînement non supervisé ou l'architecture du réseau, nous prétendons que CoNLL03 et OntoNotes évaluent mal la capacité de généralisation des algorithmes à cause d'un important recoupement lexical entre les mentions présentent dans le jeu d'entraînement et les jeux de validation et de test. Nous le quantifions en séparant les occurrences des mentions dans les sets d'évaluation en trois catégories : recoupement exact, recoupement partiel et recoupement nul, de manière similaire à Augenstein et al. [ADB17].

Une mention d'un jeu d'évaluation est un recoupement exact si elle apparaît sous l'exacte même forme sensible à la capitalisation dans le jeu d'entraînement et annotée avec le même type. Le recoupement est partiel s'il n'est pas exact mais qu'au moins un des mots non vides de la mention apparaît dans une mention de même type. Toutes les autres mentions ont un recoupement nul : leurs mots non vides ne sont jamais rencontrés pendant l'entraînement. Ainsi, la proportion de recoupements partiels et nuls reflète la capacité d'un jeu de test à évaluer la capacité de généralisation d'un algorithme aux entités non rencontrées, ce qui est un premier pas nécessaire à l'adaptation de domaine.

Comme reporté dans la Table 1, les deux jeux de données montrent un important recoupement lexical de mentions. Dans CoNLL03, plus de la moitié des occurrences de mentions du jeu de test est présente dans le jeu d'entraînement alors que seulement 28% sont totalement nouvelles. Dans OntoNotes, le recoupement est encore pire avec 67% de recoupement exact contre 9% de nouvelles mentions. De plus, nous remarquons une influence significative du type d'entité puisque LOC et MISC présentent le recoupement le plus important alors que PER et ORG ont un vocabulaire plus varié.

Cela montre que les deux principaux jeux de données étalons en REN en anglais évaluent surtout la performance d'extraction des mentions déjà rencontrées lors de l'entraînement, bien qu'apparaissant dans des phrases différentes. De telles proportions de recoupement lexical ne sont pas réalistes dans des applications réelles où un modèle doit traiter quelques ordres de grandeurs de documents de plus en inférence qu'en entraînement pour rentabiliser le coût de l'annotation. L'amélioration spécifique des performances sur les nouvelles mentions revêt donc une importance cruciale dans un cas concret qui est sous-estimée par les benchmarks actuels.

Nous proposons donc une évaluation extra domaine en entraînant les modèles sur CoNLL03 et en les testant sur OntoNotes, plus grand et plus diversifié, ce qui correspond mieux au cas concret. Nous gardons les types de CoNLL03 et y alignons ceux d'OntoNotes : ORG et PER correspondent déjà et nous alignons LOC + GPE dans OntoNotes à LOC dans CoNLL et NORP + LANGUAGE à MISC. Cela réduit le recoupement exact à 42%, ce qui nous semble encore une surestimation du recoupement en utilisation réelle.

				CoNLL03				OntoNotes			
Entraînement	Modèle	Représentation	Dim	Exact	Partiel	Nul	Tous	Exact	Partiel	Nul	Tous
CoNLL03	**BiLSTM-CRF**	BERT	4096	95.7	88.8	82.2	90.5	95.1	82.9	73.5	**85.0**
		ELMo	1024	95.9	89.2	85.8	**91.8**	94.3	79.2	72.4	83.4
		Flair	4096	95.4	88.1	83.5	90.6	94.0	76.1	62.1	79.0
		GloVe + char	350	95.3	85.5	83.1	89.9	93.9	73.9	60.4	77.9
		GloVe	300	95.1	85.3	81.1	89.3	93.7	73.0	57.4	76.9
	Map-CRF	BERT	4096	93.2	85.8	73.7	86.2	93.5	77.8	67.8	80.9
		ELMo	1024	93.7	87.2	80.1	**88.7**	93.6	79.1	69.5	**82.2**
		Flair	4096	94.3	85.1	78.6	88.1	93.2	74.0	59.6	77.5
		GloVe + char	350	93.1	80.7	69.8	84.4	91.8	69.3	55.6	74.8
		GloVe	300	92.2	77.0	61.7	81.5	89.6	62.8	38.5	68.1
OntoNotes	**BiLSTM-CRF**	BERT	4096					96.9	88.6	81.1	**93.5**
		ELMo	1024					97.1	88.0	79.9	93.4
		Flair	4096					96.7	85.8	75.0	92.1
		GloVe + char	350					96.3	83.3	69.9	91.0
		GloVe	300					96.2	82.9	63.8	90.4

TABLE 2 – Scores micro-F1 séparés par degré de recoupement en évaluation intra et extra domaine. Nos résultats sont obtenus en moyennant cinq entraînements.

4 Représentations de Mots

Plongements de mots classiques Nous prenons **GloVe** [PSM14] comme base de référence des plongements traditionnels. Bien que les plongements GloVe soient calculés sur un corpus important pour capturer une similarité sémantique basée sur la co-occurrence, cette représentation est purement lexicale puisque chaque mot est aligné à une unique représentation. Les plongements sont initialisés avec GloVe 840B et leurs valeurs sont affinées pendant l'entraînement.

Plongements de mots à l'échelle des caractères Nous reproduisons le **Char-BiLSTM** de Lample et al. [LBS+16], un BiLSTM au niveau de chaque mot qui apprend sa représentation à partir des plongements de ses caractères pour tenir compte de caractéristiques orthographiques et morphologiques. Le Char-BiLSTM est entraîné conjointement au réseau de REN et ses sorties sont concaténées aux plongements GloVe.

Plongements de mots contextuels Contrairement aux représentations précédentes, les plongements de mots contextuels prennent en compte le contexte d'un mot dans sa représentation. Pour se faire, un modèle de langue est préentrainé sur un corpus non annoté et on prend sa représentation interne de la prédiction d'un mot sachant son contexte. **ELMo** [PNI+18] utilise un réseau convolutif à l'échelle des caractères (Char-CNN) pour obtenir un plongement de mot indépendant du contexte et la concaténation de modèles de langue LSTM à deux couches en sens avant et inverse pour la contextualisation. **BERT** [DCLT18] adopte des plongements de sous-mots et apprend une représentation dépendant des contextes droits et gauches en entraînant l'encodeur d'un Transformer [VSP+17] pour un modèle de langue masqué et la prédiction de la phrase suivante. Nous utilisons le modèle "$\text{BERT}_{\text{LARGE}}$ feature-based" pour une comparaison plus juste : les poids du modèle de langue sont gelés et nous concaténons les états cachés de ses quatre dernières couches. **Flair** [ABV18] emploie directement un modèle de langue à l'échelle du caractère. Comme pour ELMo, deux modèles de langue LSTM de sens opposés sont entraînés et leurs sorties concaténées. Flair et ELMo sont pré-entraînés sur le 1 Billion Word Benchmark [CMS+13] alors que BERT l'est sur la réunion de Book Corpus [ZKZ+15] et Wikipedia en anglais.

5 Expériences

5.1 Cadre Expérimental

Pour effectuer la REN, nous plaçons les représentations de mots dans deux modèles : un BiLSTM-CRF [HXY15] avec une dimension cachée de 100 dans chaque direction et Map-CRF pour lequel elles sont projetées linéairement dans l'espace de sortie. Nous gardons le CRF [LMP01] car la projection de plongements non contextuels revient à une prédiction indépendante pour chaque mot.

Nous séparons les Précision, Rappel et score F1 par degré de recoupement exact, partiel ou nul. Pour la Précision, cette séparation est effectuée a posteriori sur les prédictions du modèle. Nous utilisons le schéma

d'annotations IOBES et validons sur le score-micro F1. Nous rapportons les moyennes des scores obtenus par 5 entraînements différents. Pour chaque modèle, nous choisissons le meilleur de SGD ou Adam avec un taux d'apprentissage de 0.001, des batchs de taille 64, un dropout de 0.5 et un early stopping avec patience 5.

Les scores F1 intra et extra domaine des modèles entraînés sur CoNLL03 sont rapportés dans la Table 2 ainsi que les bornes supérieures extra domaine obtenues en entraînement sur OntoNotes réaligné. Nous omettons délibérément l'évaluation extra domaine de OntoNotes vers CoNLL03 en considérant que les cas d'application concrets sont toujours limités en ressources annotées et ainsi entraîné sur le jeu de données le plus petit et le moins varié.

5.2 Résultats

Performances Intra Domaine Tout d'abord, dans toutes les configurations le score F1 est le plus haut pour les recoupements exacts, puis partiels et nuls ce qui confirme le biais dans les jeux de données avec un recoupement lexical important. Ensuite, bien que ELMo apparaît comme la solution la plus stable intra domaine, il est difficile de dégager une hiérarchie claire entre plongements contextuels puisque les données de pré-entraînement ainsi que la dimension des représentations diffèrent. Pour BERT et Flair, le BiLSTM-CRF performe relativement moins bien sur CoNLL03 que sur OntoNotes, probablement par surapprentissage sur CoNLL03. De plus, le gain maximal de la contextualisation sur CoNLL03 est de +0.6 F1 en recoupement exact contre +3.7 en partiel et +2.7 en nul. D'autre part, Map-CRF avec ELMo ou Flair arrive presque au même niveau que BiLSTM-CRF et Glove + char, ce qui montre que les modèles de langues capturent intrinsèquement des représentations utiles à la REN. Enfin, quelle que soit la représentation le BiLSTM réduit l'écart de performance entre mentions vues et non vues.

Généralisation Extra Domaine En évaluation extra domaine, les performances se dégradent et l'écart se creuse entre les mentions vues et non vues. De plus, la contextualisation est encore plus bénéfique aux mentions non vues avec +1.2 F1 en recoupement exact, +9.0 en partiel et +13.1 en nul avec le BiLSTM-CRF et BERT. Cette amélioration provient clairement du pré-entraînement du modèle de langue puisque même avec Map-CRF, les plongements contextuels atteignent au moins 77.5 F1 contre 77.9 pour BiLSTM-CRF et GloVe + char. Nous distinguons néanmoins une séparation entre plongements contextuels puisque Flair, issu d'un modèle de langue à l'échelle des caractères, généralise moins bien que ELMo ou BERT en extra domaine pour les deux modèles. Il ressort ainsi que contextualiser des mots ou sous-mots conduit à une meilleure généralisation en REN. Enfin, nous pouvons séparer les performances par genres des documents dans OntoNotes comme rapporté dans la Table 3. Pour tous les modèles, la meilleure adaptation se fait pour le type *broadcast news* qui est plus proche du domaine de CoNLL03 que *web text* ou *magazine*. Cependant, les plongements contextuels bénéficient principalement aux genres plus distants et mènent à des résultats plus homogènes.

	bc	bn	nw	mz	tc	wb	Tous
BERT	87.2	88.4	84.7	82.4	84.5	79.5	**85.0**
ELMo	85.0	88.6	82.9	78.1	84.0	79.9	83.4
Flair	78.0	86.5	80.4	71.1	73.5	72.1	79.0
GloVe + char	80.4	86.3	77.0	70.7	79.7	69.2	77.9

TABLE 3 – Scores micro-F1 extra domaine du BiLSTM-CRF par genres. Respectivement *broadcast conversation*, *broadcast news*, *news wire*, *magazine*, *telephone conversation* et *web text*.

Influence du Type Bien que le score micro-F1 soit souvent la seule métrique rapportée en REN, les types d'entités devraient être pris en compte. Comme montré dans la Table 4, en intra domaine MISC est le type le plus difficile à détecter alors que PER est le plus facile, certainement grâce à un motif prénom-nom fréquent. Cependant, la contextualisation bénéficie homogènement à tous les types en intra domaine alors qu'elle bénéficie surtout à ORG et PER en extra domaine. Cela s'explique par moins de 18% de recoupement exact avec le set d'entraînement contre plus de 70% pour LOC et MISC. Ainsi, la contextualisation est plus utile pour la généralisation aux types avec le plus de variation lexicale même quand ils sont plus faciles à détecter en intra domaine.

6 Travaux Connexes

Augenstein et al. [ADB17] présentent une étude quantitative de deux modèles basés sur les CRF et un réseau convolutif avec des plongements de mots classiques [CW11] sur sept jeux de données dont CoNLL03 et OntoNotes. Ils séparent notamment les performances sur les mentions rencontrées en entraînement (notre recoupement exact) de celles non rencontrées et montrent une chute du score F1 sur ces dernières.

	LOC				MISC				ORG				PER				Tous
	Exact	Partiel	Nul	Tous	Exact	Partiel	Nul	Tous	Exact	Partiel	Nul	Tous	Exact	Partiel	Nul	Tous	Tous
CoNLL03 → CoNLL03																	
BERT	96.1	68.7	76.5	92.1	93.6	53.5	46.9	79.7	95.2	82.0	79.1	88.0	99.0	98.1	93.2	96.2	90.5
ELMo	96.0	72.6	83.3	93.1	94.3	58.9	53.9	81.8	96.0	80.8	83.5	89.6	98.9	98.7	94.6	97.0	**91.8**
Flair	95.7	73.3	79.7	92.3	93.5	55.9	49.2	80.2	95.1	77.9	80.9	87.8	98.3	98.3	93.3	96.2	90.6
GloVe + char	95.6	64.1	80.5	91.8	93.3	54.0	40.8	78.9	94.9	74.4	82.0	87.5	98.7	97.2	92.0	95.2	89.9
CoNLL03 → OntoNotes																	
BERT	96.1	65.7	79.0	89.6	94.1	51.2	25.8	72.6	93.6	83.6	76.4	82.6	93.5	90.3	83.4	88.2	**85.0**
ELMo	94.9	63.0	77.7	88.5	94.6	56.5	37.8	78.8	92.8	80.8	74.5	80.5	91.9	84.5	75.6	82.2	83.4
Flair	95.3	59.7	67.8	86.2	94.0	52.8	28.4	74.2	89.3	77.2	59.9	72.6	92.0	82.1	70.1	78.8	79.0
GloVe + char	95.6	62.4	69.3	86.7	93.8	56.7	30.1	75.3	88.9	74.0	58.5	70.8	89.5	78.9	64.7	74.8	77.9

TABLE 4 – Scores F1 par type du BiLSTM-CRF entraîné sur CoNLL03 en évaluation intra et extra domaine.

Moosavi et Strube [MS17] soulèvent un phénomène similaire en Résolution de Coréférence sur CoNLL-2012 et montrent qu'en évaluation extra domaine l'écart de performance entre les modèles d'apprentissage profond et un système de règles disparaît. Dans [MS18], ils proposent d'utiliser des caractéristiques linguistiques (comme le genre, le type d'entité ou la catégorie grammaticale) pour améliorer la généralisation extra domaine. Néanmoins, de telles caractéristiques sont obtenues en utilisant des modèles à leur tour entraînés avec des caractéristiques lexicale et sur des données ou le même problème de recoupement lexical se pose, au moins pour la Reconnaissance d'Entités Nommées.

7 Conclusion

Les benchmarks actuels de REN sont donc biaisés en faveur des mentions déjà rencontrées, à l'exact opposé des applications concrètes. D'où la nécessité de séparer les performances par degré de recoupement des mentions pour mieux évaluer les capacités de généralisation. Dans ce cadre, les plongements contextuels bénéficient plus significativement aux mentions non rencontrées, d'autant plus en extra domaine.

Les travaux futurs peuvent chercher à réduire encore l'écart de performance entre mentions rencontrées ou non, améliorer les capacités d'adaptation de domaine zero-shot avec des données cibles additionnelles ou aborder la généralisation multilingue en utilisant des modèles de langues entraînés sur des corpus multilingues.

Remerciements

Nous remercions Geoffrey Scoutheeten et Victor Storchan pour leurs points de vue et commentaires précieux. Ce travail est principalement financé par BNP Paribas dans le cadre de la convention CIFRE 2018/0327.

Références

[ABV18] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[ADB17] Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition : A quantitative analysis. *Computer Speech & Language*, 44 :61–83, 7 2017.

[CMS+13] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv preprint arXiv :1312.3005*, 2013.

[CW11] Ronan Collobert and Jason Weston. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 :2493–2537, 2011.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[HR18] Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339, 2018.

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv :1508.01991*, 2015.

[LBS+16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT 2016*, pages 260–270, 2016.

[LDS18] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. Transfer Learning for Named-Entity Recognition with Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 4470–4473, 2018.

[LMP01] John Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, volume 8, pages 282–289, 2001.

[MS17] Nafise Sadat Moosavi and Michael Strube. Lexical Features in Coreference Resolution : To be Used With Caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 14–19, 2017.

[MS18] Nafise Sadat Moosavi and Michael Strube. Using Linguistic Features to Improve the Generalization Capability of Neural Coreference Resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, 2018.

[PMX+12] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task : Modeling multilingual unrestricted coreference in OntoNotes. *Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics*, pages 1–40, 2012.

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2 2018.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[RCL18] Juan Diego Rodriguez, Adam Caldwell, and Alexander Liu. Transfer Learning for Entity Recognition of Novel Classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1974–1985, 2018.

[RSN+18] Alec Radford, Tim Salimans, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. page 12, 2018.

[SDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh Conference on Natural Language Learning at NAACL-HLT 2003*, volume 4, pages 142–147, 2003.

[SVBM17] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, 2017.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[WPM+13] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, and Michelle Franchini. OntoNotes Release 5.0 LDC2013T19. *Linguistic Data Consortium, Philadelphia, PA*, 2013.

[ZKZ+15] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies : Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, 2015.