SLICE 2.0: Weakly supervised interpretable word embedding learning and evaluation

PUPIER Adrien

Aix-Marseille Université
Master 2 Informatique : IAAA

Juin 2021

Introduction

Words embeddings

- Vecteur de nombres de dimension D qui représente un mot.
- Deux familles : Non-contextuel (Mikolov et al. (2013)) et contextuel. (Vaswani et al. (2017))

1. source de l'image : https ://predictivehacks.com/a-high-level-introduction-to-word-embeddings/

Introduction

Problèmes avec les words embeddings

- Individuellement non interprétable. Pour interpréter un embedding, on doit regarder ses voisins les plus proches.
- On aimerait pouvoir connaitre le sens d'un mot uniquement à partir de son embedding.

WordNet, Supersense et Hypersense

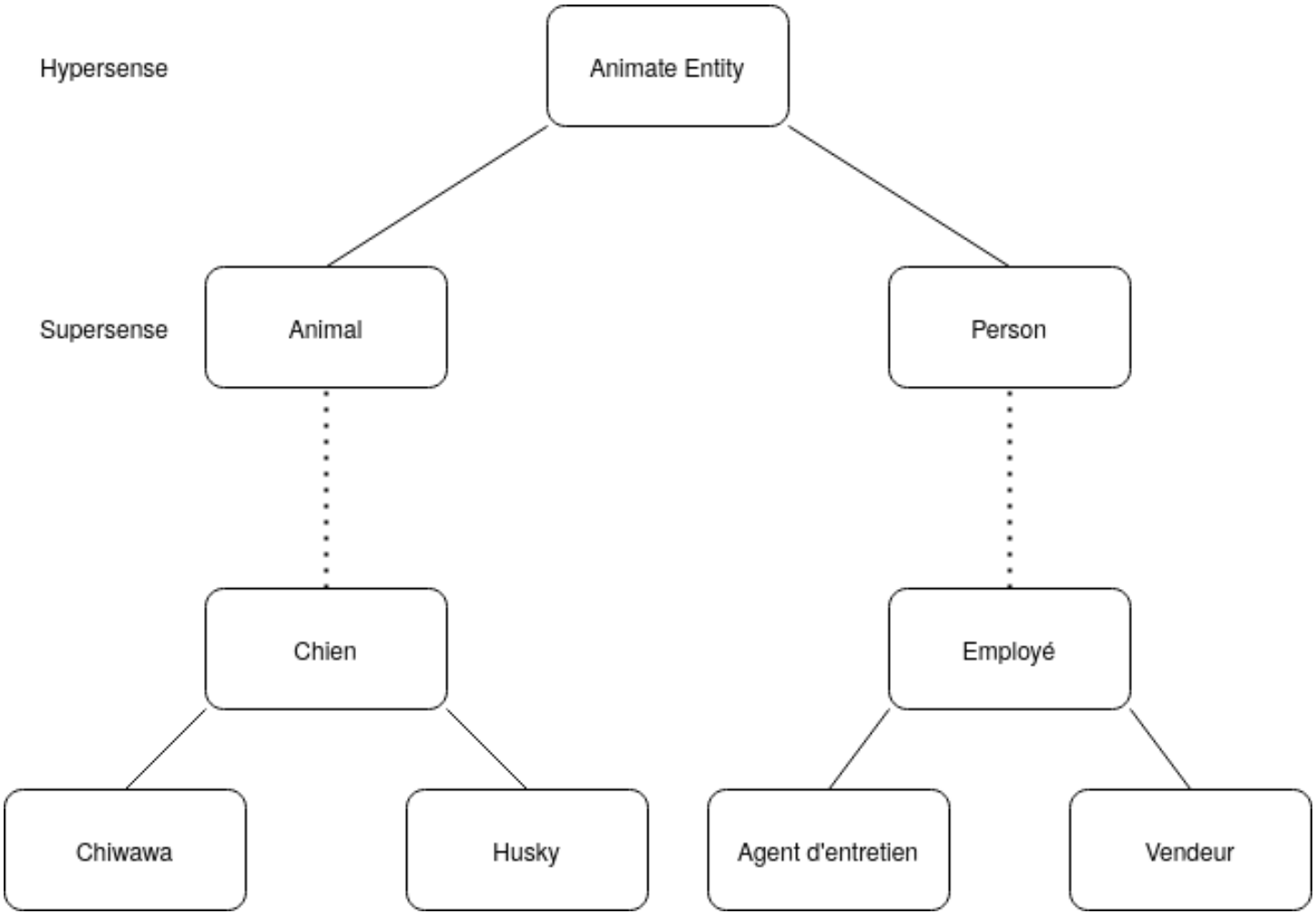


Figure – Schéma des différentes granularités des inventaires de sens.

Supersense et Hypersense

Hypersenses	**WordNet Supersenses**
Living Animate Entity (ANI)	Animal, Person
Manufactured Object (MAN)	Artifact
Natural Object (NAT)	Body, Plant, Object
Informational Object (INF)	Cognition, Communication, Possession
Dynamic Situation (DYN)	Act, Event
Stative Situation (STA)	Attribute, State, Feeling
No Hypersenses	Food, Groups, Institutions, Location, Phenomenon, Process, Quantity, Relation, Shapes, Substance, Time

SLICE 1.0

Supersense-based **L**ightweight **I**nterpretable **C**ontextual **E**mbeddings

- Modèle créée par l'équipe TALEP au laboratoire du LIS (Aloui et al. (2020))
- Objectif : Créer des words embeddings interprétables avec peu de supervision sur des noms en français.
- Dans le cas du papier SLICE, l'inventaire de sens $\Rightarrow$ Hypersenses.

SLICE embeddings

- Le modèle créé des embeddings contextuels où chaque dimension correspond à un sens précis.
- e.g : "Il a vu une (grue) manger." "Ils ont utilisé une (grue)."

ANI	MAN
0.9	0.2

ANI	MAN
0.1	0.9

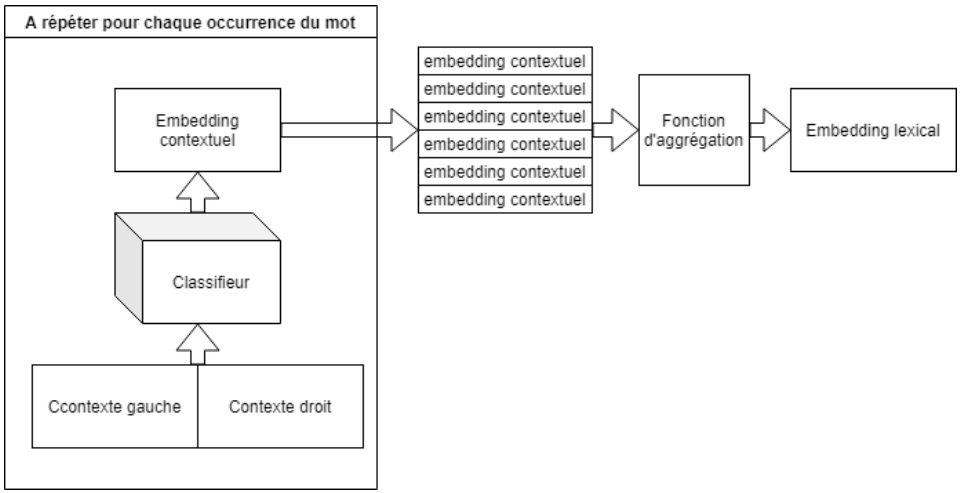


Figure – Schéma de la création des embeddings pour un mot.

Entrainement des classifieurs binaires

- Chaque sens nécessite une liste de mots monosémiques (seeds) liée à ce sens. e.g : *car* $\Rightarrow$ MANUFACTURED OBJECT. Deux listes, une positive et une négative.

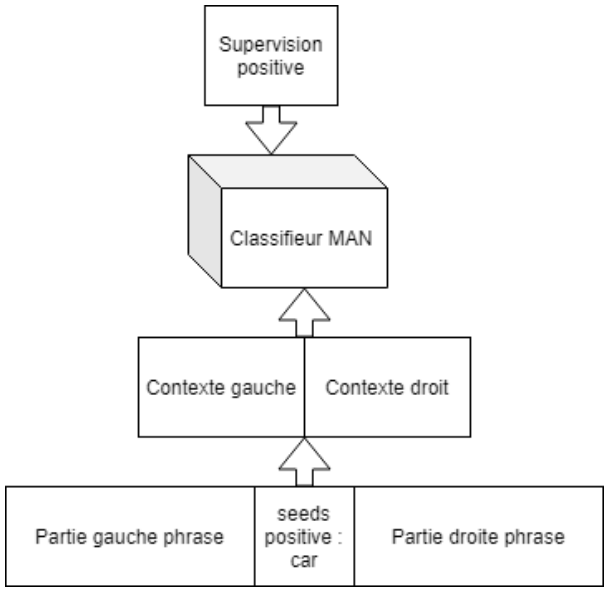


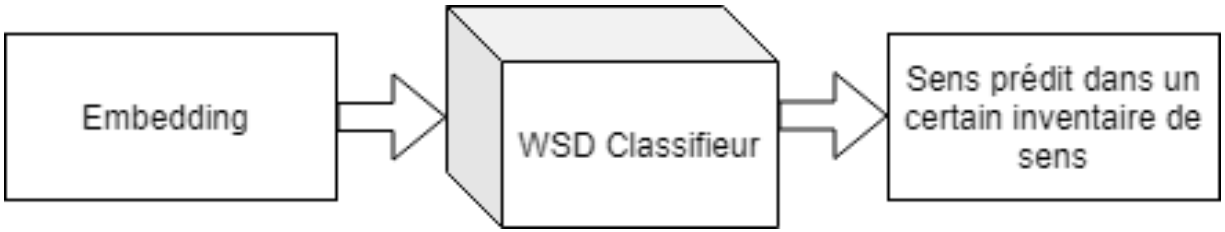
Figure – Schéma de l'entraînement d'un classifieur binaire

Question

- Est-il possible d'appliquer la méthode SLICE sur d'autres langues, d'autres inventaires de sens en essayant de la rendre la plus légère possible ?

Word sense disambiguation

- Pour évaluer les embeddings, une tâche de Word Sense Disambiguation (WSD) est mise en place.
- Pour cela, nous utiliserons un MLP.
- Les corpus utilisés sont : le SemCor (Miller et al. (1993)) pour l'anglais et le FR-SemCor (Barque et al. (2020)) pour le français.
- Tous les résultats suivants sont obtenus sur cette tache.



Q1 : Modèle Multi-class

	Accuracy
Binaire	$0.829 \pm 0.005^*$
Multi-class	$0.835 \pm 0.005^*$

Table – Accuracy sur la tâche de WSD sur le FR-SemCor avec les Hypersenses. P-value = 0.0152

- Résultat supérieur au binaire sur notre tâche.
- Gain de temps conséquent. Un seul classifieur à entraîner contre N avec SLICE 1.0.
- Changement de philosophie : les sens sont maintenant en concurrence alors qu'ils étaient indépendants dans la version 1.0.

Résultats sur l'Anglais et le Français avec Hypersense et Supersense

Français	Accuracy
Hypersense	0.835 ± 0.006
Supersense	0.833 ± 0.005

Anglais	Accuracy
Hypersense	0.750 ± 0.004
Supersense	0.685 ± 0.002

Résultats sur l'Anglais et le Français avec Hypersense et Supersense

	Français	Anglais
Hypersense	1.227	1.490
Supersense	1.276	1.673

Table – Difficulté des différents corpus en termes de polysémie moyenne (PLM)

Commentaire

- Pas de différence élevée entre les résultats Français sur les différentes granularités.
- Les résultats sont corrélés à la polysémie moyenne du corpus.

Exploiter les differentes granularités

Les differentes methodes

- Les Supersenses et Hypersenses ont une relation hiérarchique. Peut-on utiliser les Supersenses pour aider la prédiction de Hypersenses ? L'inverse est-il possible ?
- 4 Méthodes testées.

M1 : Remplacer le Supersense prédit par son Hypersense

- Ajout d'un Hypersense OTHERS
- Map le Supersense vers son Hypersense

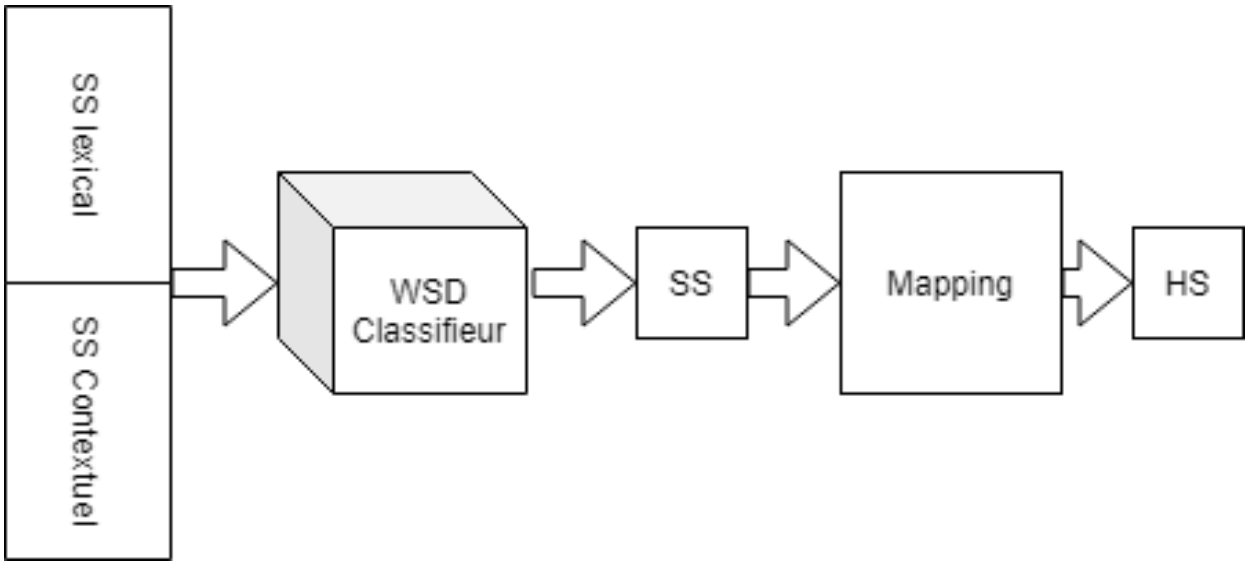


Figure – Schéma de la méthode de mapping

M2 : Utiliser l'embedding Supersense pour prédire l'Hypersense

- On remplace la cible Supersense par son Hypersense dans les données du classifieur.

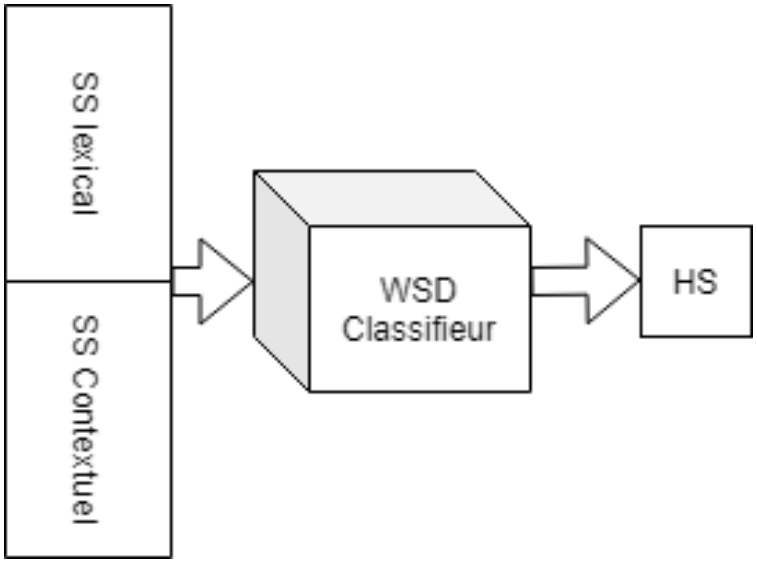


Figure – Schéma de la méthode de prédiction de Hypersense avec embedding Supersense

M3 : Concaténer l'embedding Supersense et Hypersense.

- On donne au classifieur, pour un même mot, son embedding en Hypersense et en Supersense

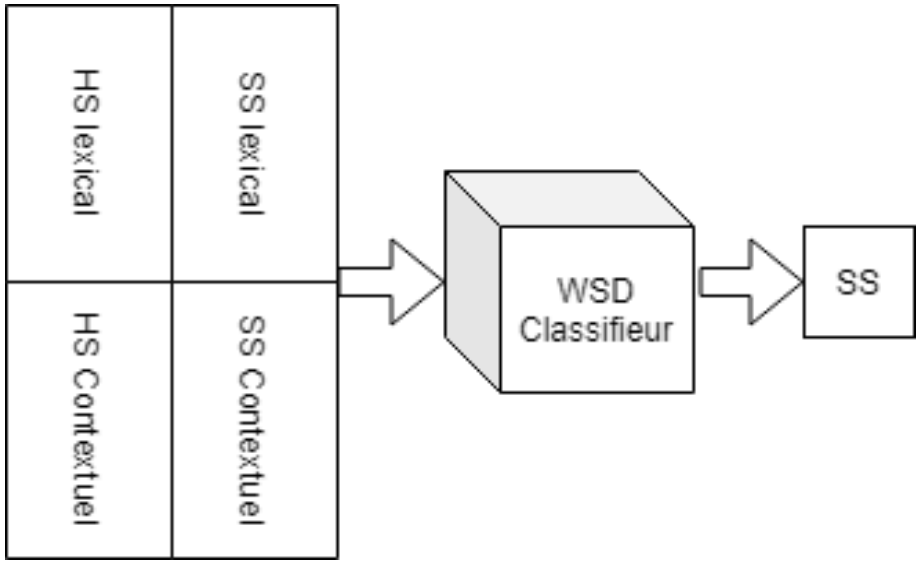


Figure – Schéma de la méthode de concatenation d'embedding

M4 : Prédire simultanément l'Hypersense et le Supersense.

- Classifieur multitâche (Caruana (1997))
- Défaut : Prise de décision indépendante pour chaque tâche $\Rightarrow$ apparition de décision incohérente.

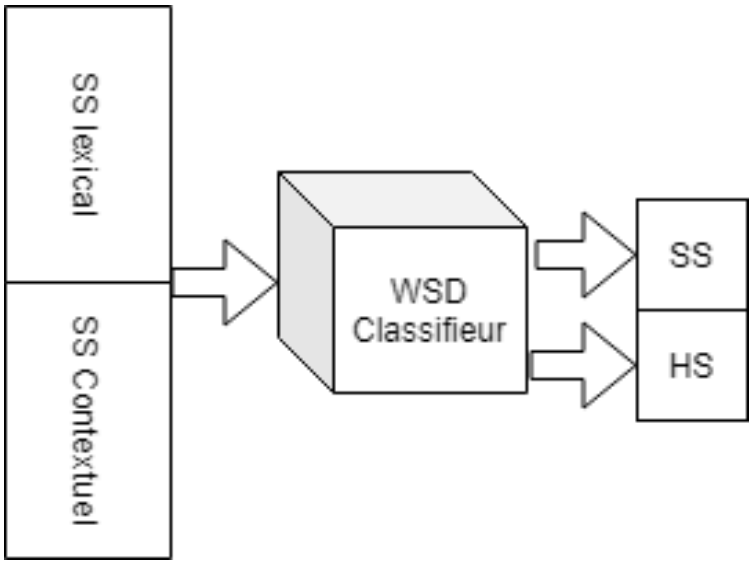


Figure – Schéma de la méthode multi-tâche

Résultats

Méthodes	Anglais HS	Anglais SS	Français HS
M1	-0.004		+0.029
M2	-0.029		+0.015
M3		-0.006	
M4	+0.004	+0.006	

Table – Différence entre la moyenne de la méthode originale et les différentes méthodes dans cette partie.

- Dans tous les cas, les résultats sur l'anglais reste très proche
- La méthode 4 (multi-tâche) est la plus prometteuse.
- La catégorie OTHERS n'est pas satisfaisante.

Impact de la tailles des listes de graines sur les embeddings

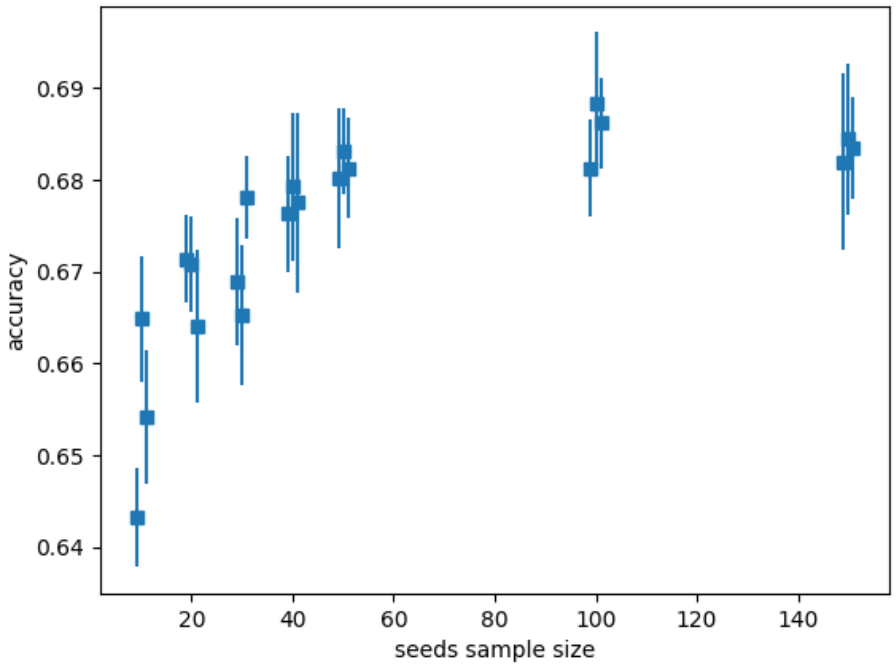


Figure – Accuracy en fonction de la taille des listes de seeds sur l'Anglais.

- Même avec peu de seeds (10), les résultats restent corrects.
- La valeur optimale semble être entre 50 et 100 seeds par sens.
- Bonne nouvelle pour appliquer cette technique à d'autres langues.

Conclusion

En résumé :

- Les embeddings produits par le classifieur multi-class sont meilleurs sur notre tâche d'évaluation.
- Il est possible d'utiliser cette méthode sur une granularité plus fine avec un classifieur multi-class.
- Cette méthode est applicable à d'autres langues comme l'Anglais.
- Le multi-tâche est prometteur dans le cadre de la tâche de WSD.
- Le nombre de seeds nécessaire est bien plus faible que ce qui est utilisé dans SLICE 1.0

Perspective de travail futur

- Impact de la température du softmax (Hinton et al. (2015)) sur les embeddings produits.
- Ajout de couche d'attention/self attention dans le modèle.
- Réduire les incohérences du classifieur multitâche sur la tâche de WSD.
- Utiliser un autre inventaire de sens que les Supersenses de WordNet comme CSI (Lacerra et al. (2020))

Annexe M1

	Accuracy
Français Hypersense mapped	$0.864 \pm 0.002^*$
Français Hypersense Original	$0.835 \pm 0.006^*$
Anglais Hypersense mapped	$0.746 \pm 0.003^*$
Anglais Hypersense Original	$0.750 \pm 0.004^*$

Table – Accuracy quand on fait correspondre le Supersense à son Hypersense. Français p-value < 0.0001. Anglais p-value=0.0210

Annexe M2

	Accuracy
Français sans others	0.867 ± 0.004
Français avec others	0.850 ± 0.003
Français Hypersense Original	0.835 ± 0.006
Anglais sans others	0.755 ± 0.004
Anglais avec others	0.721 ± 0.004
Anglais Hypersense Original	0.750 ± 0.004

Table – Résultats des prédictions Hypersenses en utilisant les embeddings Supersenses

Annexe M3

	Accuracy
Anglais concaténé	0.679 ± 0.003
Anglais original	0.685 ± 0.002

Table – Accuracy avec des embeddings de taille 60 (24+24 Supersenses, 6+6 Hypersenses)

Annexe M4

	Accuracy
Anglais Hypersense	$0.754 \pm 0.002^*$
Anglais Hypersense Original	$0.750 \pm 0.004^*$
Anglais Supersense	$0.691 \pm 0.003^*$
Anglais Supersense Original	$0.685 \pm 0.002^*$

Table – Accuracy sur l'Anglais avec le modèle multitâche sur les Hypersenses et Supersenses. Hypersense p-value= 0.0111, Supersense p-value < 0.0001

Courbe d'apprentissage

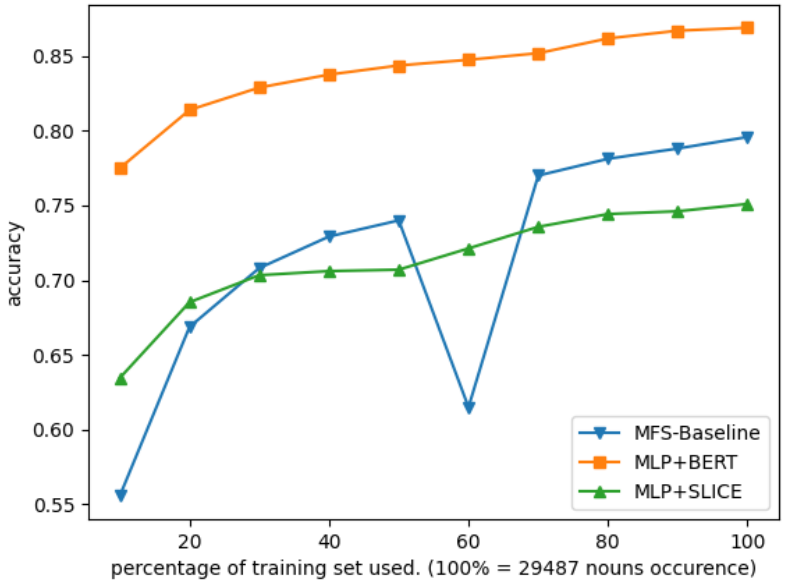


Figure – Courbe d'apprentissage sur l'anglais avec la granularité Hypersense.

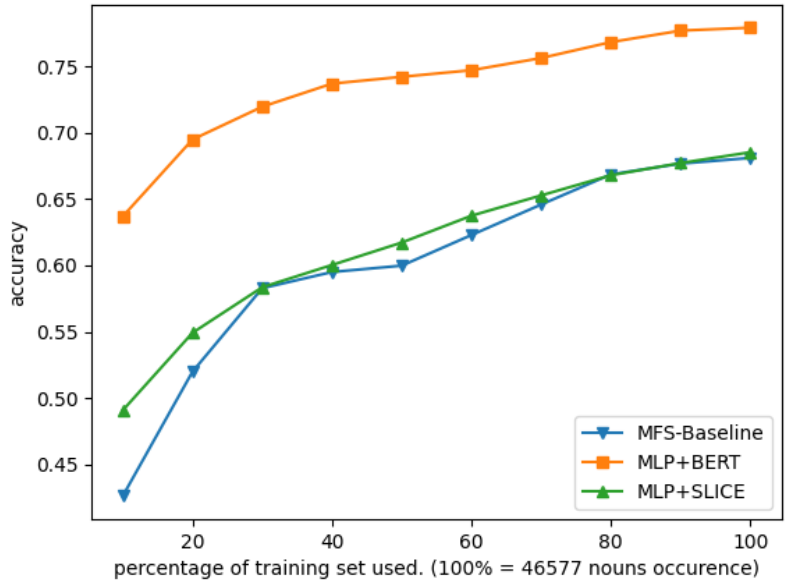


Figure – Courbe d'apprentissage sur l'anglais avec la granularité Supersense.

References I

Aloui, C., Ramisch, C., Nasr, A., and Barque, L. (2020). Slice : Supersense-based lightweight interpretable contextual embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370.

Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., and Segonne, V. (2020). Frsemcor : Annotating a french corpus with supersenses. In *LREC-2020*.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1) :41–75.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*.

Lacerra, C., Bevilacqua, M., Pasini, T., and Navigli, R. (2020). Csi : A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8123–8130.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv :1310.4546*.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY : Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv :1706.03762*.