TD 4 - Analyses

Carlos Ramisch et Manon Scholivet - Recherche Zen

2023-2024

1 Jupyter notebook

Nous avons préparé quelques exercices à faire pour mieux comprendre les notions du cours. Ces exercices sont donnés sous la forme d'un Jupyter Notebook où vous pourrez tester avec du code en python les notions vues en cours. Ces exercices utilisent les données issues de l'annotation de compositionnalité (cf. TD précédent).

`https://colab.research.google.com/drive/1HzLh4657TJbh_1vKQT6YKfvoXKniXgJO?usp=sharing`

On vous recommande de faire une copie de ce fichier pour pouvoir le modifier et travailler librement. Vous pouvez aussi le télécharger pour travailler sur votre ordinateur si vous avez Jupyter notebook installé sur votre ordinateur.

2 Étiquetage multilingue

Avoir de nombreux résultats après avoir fait de nombreuses expériences, c'est super! C'est maintenant le moment d'aller présenter vos résultats à vos collègues. Cependant... Si vous leur présenter d'immenses tableaux de résultats, il est plus que probable qu'une personne vous dise : "Mais du coup... Qu'est-ce que tu conclues de tous ces résultats ?". Si vous avez parfaitement défini votre question de recherche et votre protocole expérimental, les conclusions peuvent être "évidentes". Mais il arrive que certaines configurations n'ont pas été anticipée, ou que l'on n'est pas assez poussé les réflexions... Nous allons donc essayer d'analyser un gros tableau de résultat afin d'essayer d'en tirer tout de même quelques informations intéressantes.

Quelles sont les informations de ce tableau ? Tout d'abord, tous les résultats présentés correspondent aux résultats de modèles entraînés pour résoudre une tâche de prédiction : celle des parties du discours (*Part Of Speech (POS)*. La parties du discours d'un mot est tout simplement sa catégorie : est-ce un nom ? un verbe ? un adjectif ?

La question générale de recherche était (en gros) la suivante : l'utilisation de ressources typologiques permet elle de mieux prédire les POS lors de l'absence de données d'entraînement ?

- Les expériences *Mono* correspondent à la *baseline*, voir une *topline*. Ces systèmes ne sont pas multilingues mais monolingues. Si on prend la première ligne du tableau, correspondant à l'arabe (ar) : on a entraîné un modèle sur de l'arabe, et on test sur des données... en arabe ! Tout simplement. Pour chacune des deux colonnes, 38 modèles ont été entraînés, un pour chaque langue.[1] Nous verrons plus loin la différence entre les deux systèmes.

- Les expériences *Multi* sont multilingues. Si on reprend la première ligne, un modèle a été appris sur les 38 langues, et on teste sur l'arabe. Pour la deuxième ligne (et toutes les autres) on reprend le même modèle, et on teste sur un jeu de donnée en bulgare (bg). Un seul modèle a été entraîné pour chaque colonne.

- Les expériences *ZS* sont des expériences dite de *zero-shot*. Ces expériences permettent de simuler la situation où aucune donnée d'entraînement n'est disponible. Pour la première ligne : le modèle n'a jamais vu d'arabe à l'entraînement, mais il a eu accès aux données de 37 autres langues. Cependant, le jeu de données de test ne comprendra que des données en arabe. Pour chaque colonne, 38 modèles ont été entraînés.

[1] Les codes des langues peut être trouvé ici : `https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes`

- Les expériences $-c$ (respectivement $+c$) correspondent à l'absence (la présence) des caractères de la langue lors des expériences. Si deux langues utilisent des alphabets différents, le système ne sera donc pas capable de les distinguer.
- Les expériences ID correspondent à l'utilisation d'un identifiant de la langue dans les configurations *Multi*.
- Les expériences W_{22} correspondent à l'utilisation d'un vecteur d'informations sur la typologie des langues (les mots suivent ils un ordre Sujet-Verbe-Objet ou bien Sujet-Objet-Verbe, ...)

Le tableau des résultats sous forme de tableur est disponible ici : `https://docs.google.com/spreadsheets/d/1OyDo9UdSD_avaE87a1qEYfM4Gn2zw7IgTYAwRB9LRFc/edit?usp=sharing`

Vous trouvez ces questions ultra fun et vous voulez en savoir plus ? N'hésitez pas à consulter mon incroyable manuscrit de thèse : `https://www.theses.fr/2021AIXM0306`.

3 Prédiction de compositionnalité

Les noms composés tels que *panne sèche* ont un certain degré de **compositionnalité**, c'est-à-dire, à quel point le sens du tout (*panne sèche*) peut être vu comme une combinaison des sens des parties (une *panne* qui est littéralement *sèche*). Des annotateur.ice.s humain.e.s ont annoté chaque nom composé contenu dans des datasets en anglais (en), français (fr) et portugais (pt) avec leurs scores de compositionnalité sur une échelle allant de 0 à 5.

Ces données annotées avec leurs score de compositionnalité ont été utilisées dans la thèse de Silvio Cordeiro notamment pour des expériences décrites dans cet article : `https://aclanthology.org/J19-1001/`. Cependant, avant d'arriver à cet article, un nombre invraisemblable d'expériences ont été faites (26,304 pour être précis).

L'idée de ces expériences était de prédire automatiquement les scores de compositionnalité à l'aide de modèles de *word embeddings* tels que word2vec et Glove. Le tableau `tabelao-2017-07-21_2358.txt.tsv` contient les résultats de ces expériences, où les colonnes sont à interpréter comme suit :

- `weight` Le schéma de poids associé au mot 1 (nom) et au mot 2 (modifiuer/adjectif) pour combiner les embeddings
- `language` anglais (en), français (fr) ou portugais (pt)
- `dataset` le nom du dataset sur lequel l'expérience a été faite `model` et `submodel` modèle de word embeddings utilisé pour faire la prédiction
- `dimensions`, `lemmatype` et `window` sont trois hyper-paramètres des modèles de word embeddings
- `pearson` `spearman` `kendall` `bf1` `avgprec` `prec@` et `ndcg` sont les métriques d'évaluation qu'on voudrait maximiser. Vous pouvez vous concentrer sur `spearman`
- `misssingle` et `misscompound` dénotent le nombre de composants et composés pour lesquels aucun embeddings n'a été généré

Plusieurs questions se posent sur ces résultats, par exemple :

- Quel est le meilleur modèle et sous-modèle pour cette tâche ?
- Quelles valeurs d'hyper-paramètres marchent mieux pour chaque langue ? Pour chaque dataset ? Pour chaque modèle ?
- Est-ce que les métriques d'évaluation sont plutôt rédondantes ou complémentaires ?
- ...

Vous pouvez vous concentrer uniquement sur une partie du tableau afin de pouvoir visualiser et manipuler les résultats. Par exemple, considérer uniquement une langue, ou un dataset, ou un type de modèle. À vous de jouer pour découvrir des tendances, tirer des conclusions, voire observer des choses qui n'ont jamais été observées dans cette masse de résultats !

Lang.	*Mono* $+c$		*Mono* $-c$		*Multi* $+c$		*Multi* $+c$ W_{22}		*Multi* $+c$ ID		*Multi* $-c$		*Multi* $-c$ W_{22}		*Multi* $-c$ ID		*ZS* $+c$		*ZS* $-c$		*ZS* $+c$ W_{22}		*ZS* $-c$ W_{22}	
	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev
ar	92.94	0.10	90.80	0.05	92.03	0.23	92.56	0.02	92.57	0.05	89.09	0.02	90.35	0.13	90.38	0.17	60.23	0.58	66.88	1.73	64.94	0.85	62.42	1.41
bg	95.73	0.08	92.25	0.09	95.09	0.04	95.66	0.11	95.77	0.02	90.84	0.05	91.78	0.04	92.00	0.02	78.16	0.38	79.65	0.08	76.40	0.88	77.80	0.11
ca	95.84	0.05	90.87	0.10	95.22	0.04	96.06	0.01	96.11	0.06	89.60	0.05	91.46	0.05	91.44	0.05	83.48	0.06	77.39	0.17	82.63	0.60	76.61	0.08
cs	95.12	0.07	90.71	0.05	94.02	0.01	94.81	0.01	94.89	0.05	88.97	0.08	90.41	0.07	90.64	0.07	81.11	0.09	79.03	0.10	76.97	2.27	76.06	1.48
da	92.41	0.05	87.50	0.13	91.34	0.10	92.59	0.05	93.12	0.08	87.27	0.14	88.86	0.25	89.25	0.16	83.93	0.05	80.88	0.26	84.43	0.16	80.89	0.02
de	89.16	0.04	81.91	0.57	88.94	0.15	89.13	0.05	89.53	0.09	82.06	0.02	82.46	0.22	83.01	0.05	53.84	1.59	53.39	0.54	48.63	0.26	49.37	0.98
el	95.47	0.05	91.56	0.17	95.28	0.04	95.48	0.05	95.31	0.01	90.96	0.08	91.62	0.22	91.38	0.16	47.61	0.45	60.02	0.21	52.34	1.69	60.79	0.02
en	89.54	0.12	85.81	0.16	88.53	0.00	89.14	0.12	89.24	0.14	84.89	0.02	85.89	0.02	86.01	0.15	27.96	0.11	25.78	0.78	29.90	6.20	27.51	1.38
es	93.22	0.02	90.11	0.10	94.00	0.01	92.13	0.07	92.16	0.03	90.83	0.01	89.88	0.07	89.66	0.04	88.69	0.12	85.81	0.44	87.11	0.16	85.62	0.14
et	90.02	0.13	81.50	0.10	86.49	0.16	88.77	0.04	88.78	0.12	78.15	0.24	81.80	0.19	81.64	0.17	63.55	0.08	62.24	0.31	64.74	0.11	62.82	0.35
eu	90.62	0.10	84.73	0.33	88.16	0.11	89.92	0.05	89.75	0.02	81.16	0.11	83.94	0.08	83.94	0.05	55.73	0.11	54.87	0.20	54.36	0.38	50.77	0.61
fa	94.83	0.02	92.47	0.14	93.78	0.00	94.53	0.08	94.47	0.19	90.44	0.03	92.01	0.05	91.99	0.03	56.24	1.00	68.72	0.70	62.24	4.23	62.30	1.30
fi	84.90	0.10	77.45	0.15	82.25	0.02	83.12	0.07	83.25	0.12	75.53	0.17	77.14	0.03	77.12	0.22	67.08	0.03	65.89	0.15	66.38	0.42	64.89	0.45
fr	93.45	0.02	88.67	0.09	92.97	0.02	93.35	0.06	93.59	0.04	88.63	0.15	89.40	0.03	89.49	0.02	78.29	1.07	77.12	0.90	75.81	0.44	69.78	3.72
ga	89.11	0.09	82.84	0.12	87.02	0.03	89.09	0.23	88.78	0.12	80.23	0.14	83.30	0.17	83.16	0.11	45.28	0.32	44.20	0.36	41.03	2.33	39.20	0.48
he	94.45	0.05	92.57	0.03	93.84	0.03	94.15	0.02	94.10	0.05	91.05	0.12	92.25	0.07	92.00	0.12	52.30	0.80	55.16	0.70	53.95	2.64	55.31	0.80
hi	93.62	0.08	92.30	0.01	92.75	0.15	92.92	0.08	93.05	0.00	90.39	0.20	91.19	0.05	91.42	0.02	35.67	2.82	60.55	0.23	39.12	0.00	58.09	4.14
hr	94.30	0.07	89.42	0.19	93.38	0.12	94.28	0.10	94.22	0.07	88.00	0.01	89.66	0.09	89.78	0.25	83.07	0.16	79.01	0.59	81.79	0.81	77.98	0.58
hu	92.90	0.01	84.85	0.18	90.72	0.27	92.37	0.07	92.68	0.15	82.92	0.28	85.69	0.16	85.72	0.16	60.67	1.02	59.59	0.83	60.88	1.45	62.59	2.09
id	90.64	0.04	85.61	0.05	89.30	0.11	90.76	0.18	90.76	0.17	83.00	0.01	86.01	0.08	85.92	0.03	59.04	0.54	57.54	0.30	57.50	0.11	52.88	1.70
it	94.30	0.03	91.30	0.15	93.94	0.02	94.50	0.03	94.79	0.01	90.50	0.13	91.47	0.02	91.67	0.02	77.09	0.45	78.26	0.35	76.56	0.92	77.04	0.29
ja	92.41	0.17	88.05	0.19	91.78	0.15	92.44	0.04	92.18	0.08	87.88	0.02	88.18	0.18	88.12	0.07	33.28	1.64	45.72	2.52	37.66	1.20	40.25	0.42
ko	93.38	0.03	83.64	0.20	91.74	0.12	92.23	0.11	91.94	0.11	80.60	0.07	82.23	0.11	82.53	0.16	50.23	0.37	53.06	0.12	47.99	0.81	50.87	0.33
lv	89.44	0.15	82.25	0.06	85.84	0.12	88.69	0.12	88.84	0.21	78.53	0.05	82.69	0.27	82.70	0.11	62.75	0.10	61.31	0.12	64.46	0.23	60.21	0.09
nl	87.78	0.17	81.78	0.02	86.64	0.05	86.08	0.12	86.34	0.26	80.64	0.12	81.16	0.10	81.78	0.21	61.04	0.10	58.47	0.71	56.91	0.38	58.09	0.51
nno	90.43	0.09	84.11	0.20	90.00	0.02	83.62	0.27	83.78	1.07	84.32	0.09	78.33	0.18	70.25	3.98	84.77	0.03	79.45	0.50	79.82	0.27	75.15	0.05
nob	91.92	0.03	85.72	0.07	91.41	0.03	86.49	0.32	84.84	1.77	85.96	0.05	81.54	0.22	70.87	4.16	88.65	0.20	83.33	0.05	84.84	0.20	79.95	0.23
pl	95.19	0.00	89.12	0.28	93.46	0.11	94.41	0.03	94.55	0.08	87.09	0.10	89.27	0.12	89.16	0.05	78.82	0.16	76.08	0.08	74.30	0.22	72.78	1.76
pt	93.25	0.03	89.19	0.12	93.00	0.12	92.39	0.00	92.39	0.05	88.87	0.07	88.75	0.01	88.70	0.09	76.11	0.31	77.44	0.37	75.80	0.16	76.92	0.94
ro	94.11	0.16	89.60	0.01	92.23	0.06	93.81	0.10	93.69	0.10	86.43	0.04	89.42	0.11	89.27	0.22	64.28	0.77	65.48	0.09	65.26	0.05	65.22	1.17
ru	94.91	0.03	89.69	0.07	94.23	0.02	92.32	0.20	92.08	0.13	88.47	0.02	86.73	0.11	86.16	0.10	85.17	0.12	82.30	0.17	84.76	0.41	82.10	0.46
sl	92.50	0.05	85.06	0.09	89.48	0.01	90.13	0.30	90.79	0.01	81.64	0.20	82.95	0.14	82.88	0.20	69.10	0.29	65.31	0.07	67.29	0.40	64.44	0.62
sv	92.03	0.17	87.83	0.04	91.44	0.01	92.73	0.16	93.37	0.09	87.17	0.06	88.69	0.02	89.35	0.19	79.70	0.49	76.56	0.38	79.69	0.26	76.53	0.24
tr	90.44	0.22	80.93	0.02	86.70	0.16	88.98	0.08	88.81	0.01	77.80	0.18	81.14	0.24	81.37	0.04	56.34	0.35	55.69	0.12	54.05	0.76	49.70	1.72
uk	91.93	0.02	84.09	0.10	91.44	0.00	92.29	0.10	92.27	0.05	85.18	0.03	86.56	0.04	86.32	0.27	80.07	0.08	78.12	0.15	78.91	0.07	76.94	0.24
ur	90.56	0.01	88.78	0.29	89.53	0.10	90.51	0.12	90.44	0.08	87.84	0.14	88.70	0.03	89.02	0.09	38.97	0.21	56.66	0.51	52.82	0.03	64.95	0.33
vi	86.05	0.19	77.43	0.15	84.52	0.04	85.42	0.14	85.43	0.13	76.84	0.16	77.72	0.13	77.75	0.02	39.25	0.50	39.13	2.42	37.68	0.45	30.03	0.03
zh	87.98	0.03	82.51	0.08	88.16	0.29	88.88	0.16	88.75	0.03	83.98	0.19	84.75	0.07	84.72	0.15	50.51	0.49	49.74	0.06	49.66	0.77	52.14	0.36
Moy.	92.02	0.08	86.71	0.13	90.81	0.08	91.23	0.10	91.25	0.15	85.36	0.10	86.46	0.11	86.02	0.32	64.16	0.47	65.15	0.48	63.94	0.88	63.34	0.83

Table 1: Exactitude de la prédiction de Parties du discours de chaque système pour chaque langue