

Interprétation des mesures de similarité entre représentations neuronales de textes

Julie Tytgat^{*,**}, Guillaume Wisniewski^{*}, Adrien Betrancourt^{**}

^{*}Université Paris Cité, LLF, CNRS 75 013 Paris, France
julie.tytgat@etu.u-paris.fr & guillaume.wisniewski@u-paris.fr,

^{**}IPSIDE, 31 100 Toulouse
a.betrancourt@ipside.com

Les réseaux neuronaux utilisent des vecteurs pour représenter le texte, permettant d'évaluer leur similarité. Leur mise en oeuvre est facilitée par les nombreux modèles de langue pré-entraînés, mais malgré leurs performances dans diverses tâches, y compris la génération de textes cohérents et corrects sur le plan syntaxique, ces réseaux demeurent des "boîtes noires" dont les critères de similarité et les informations encodées ne sont pas explicitement définis.

Notre travail s'attaque à un problème très peu étudié : l'interprétation des mesures de similarité basées sur des représentations neuronales. Plus précisément, nous cherchons à déterminer si la similarité entre des représentations neuronales de texte est principalement fondée sur des critères sémantiques ou des critères de surface.

Nous proposons de mesurer, sur un corpus de 300 phrases, la similarité entre une phrase et une version de celle-ci dans laquelle un mot a été remplacé par un synonyme, un antonyme, un paronyme (prononciation similaire, sens différent) ou un synonyme du paronyme. Nous mesurons ensuite la similarité entre la phrase originale et la phrase modifiée en utilisant la similarité cosinus ou la distance euclidienne. Le tableau 1 donne un exemple issu de notre corpus.

originale	La fragrance du vol a stupéfait les gardiens du musée.
paronyme	La fragrance du vol a stupéfait les gardiens du musée.
synonyme	L' évidence du vol a stupéfait les gardiens du musée.
synonyme du paronyme	Le parfum du vol a stupéfait les gardiens du musée.
antonyme	La discretion du vol a stupéfait les gardiens du musée.

TAB. 1 – Exemples d'une phrase et de ses variantes issues de notre corpus.

Nous considérons trois modèles de langue pré-entraînés pour construire les représentations vectorielles : mBERT (Devlin et al., 2019), sBERT (Reimers et Gurevych, 2019) et le modèle d'OpenAI ADA. La figure 1a) montre la distribution des similarités cosinus entre la phrase originale et ses variantes. Les résultats montrent que la distribution des similarités et le domaine dans lequel elles varient sont très différentes d'un modèle à l'autre. Nous observons également un score de similarité moyen plus élevé pour les paires dont le niveau de surface est proche que pour celles avec un sens proche de