# RESEARCH STAY WEEK 7, Word embeddings

Daniel Cajas A01708637

# CONTEXT

For a long time, word embedding was simply seen as one more step during NLP tasks. It isn't until fairly recently, arguably 2008 as mentioned by [1] that it was split into its own task by the community. Since then there have been a lot of research into different ways of doing word embeddings and how they can be either directly used or its effects on downstream tasks.

# SEARCH METHODOLOGY

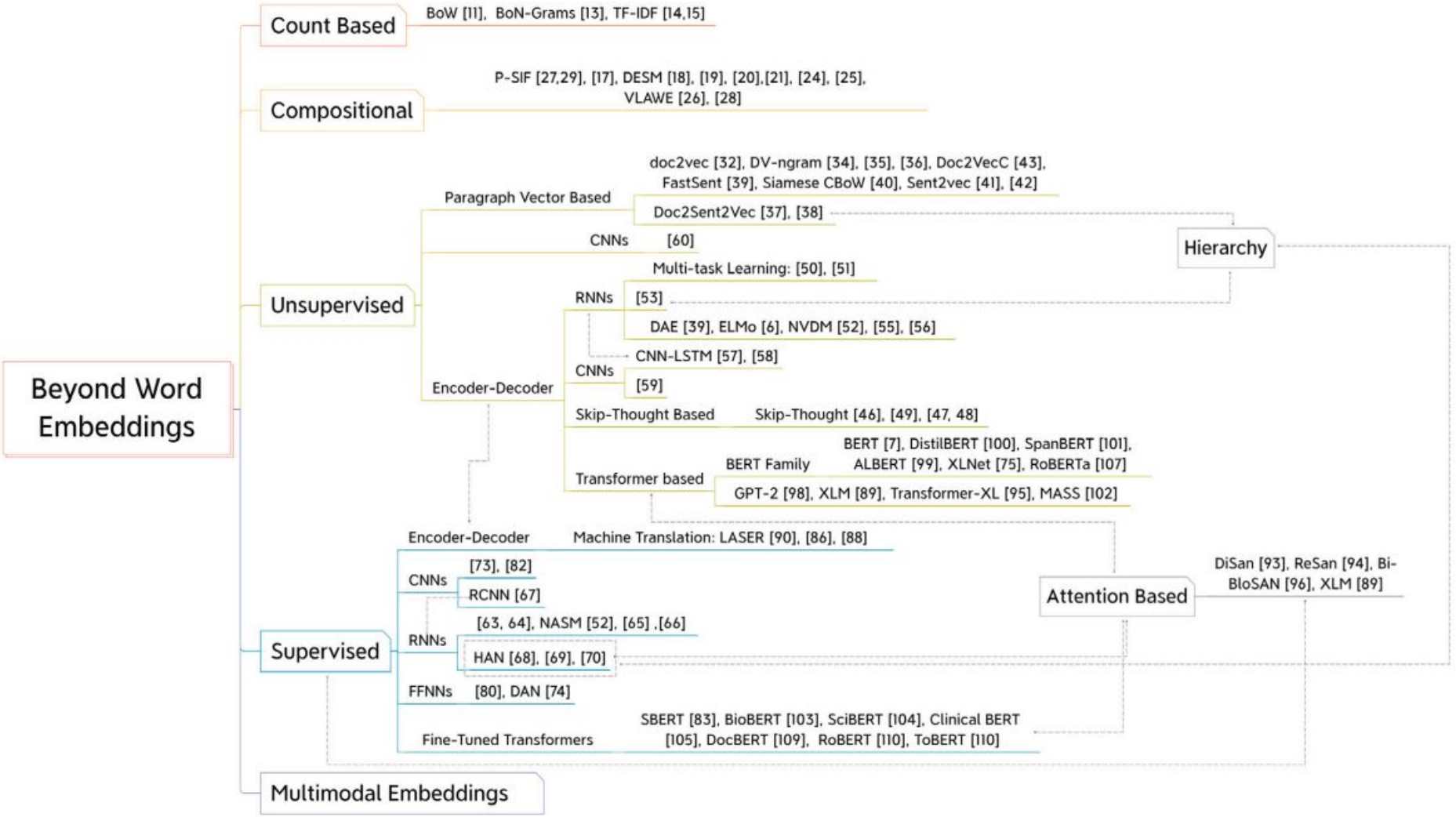
The material given for this week was both on word embeddings themselves and evaluation methods for them. I wanted to delve deeper into word embeddings. I looked in escorpus for “word embedding” surveys and found “Beyond word embeddings: A survey” [2]

# COMPARISON

| | Types of techniques | Length (Pages excluding references) | Cited | Year |
|---|---|---|---|---|
| Word Embeddings: A Survey | Prediction based, Count based | 8 | 380 | 2019 (revised in 2023) |
| Beyond Word Embeddings: A Survey | Count Based, Compositional, Unsupervised, Supervised, Transformers, Multimodal | 15 | 25 | 2023 |

# COMPARISON

Although the first paper was revised in 2023. No significant changes were made, so it is missing some critical new developments in the field. It also goes through the developments in each method chronologically in detail but doesn't do a deep dive in the state of the art and technical details. On the other hand, Beyond word embeddings [2] provides a more granular and detailed classification and description of word embedding methods, it doesn't go through the historical record of how these techniques came to be in as much detail but in exchange it explores much wider array of methods and classifies them, with very useful illustrations.


Beyond Word Embeddings
Count Based
BoW [11], BoN-Grams [13], TF-IDF [14,15]
Compositional
P-SIF [27,29], [17], DESM [18], [19], [20],[21], [24], [25], VLAWE [26], [28]
Unsupervised
Paragraph Vector Based
doc2vec [32], DV-ngram [34], [35], [36], Doc2VecC [43], FastSent [39], Siamese CBoW [40], Sent2vec [41], [42]
Doc2Sent2Vec [37], [38]
CNNs
[60]
Encoder-Decoder
Multi-task Learning: [50], [51]
RNNs
[53]
DAE [39], ELMo [6], NVDM [52], [55], [56]
CNN-LSTM [57], [58]
CNNs
[59]
Skip-Thought Based
Skip-Thought [46], [49], [47, 48]
Transformer based
BERT Family
BERT [7], DistilBERT [100], SpanBERT [101], ALBERT [99], XLNet [75], RoBERTa [107]
GPT-2 [98], XLM [89], Transformer-XL [95], MASS [102]
Hierarchy
Supervised
Encoder-Decoder
Machine Translation: LASER [90], [86], [88]
CNNs
[73], [82]
RCNN [67]
RNNs
[63, 64], NASM [52], [65] ,[66]
HAN [68], [69], [70]
FFNNs
[80], DAN [74]
Fine-Tuned Transformers
SBERT [83], BioBERT [103], SciBERT [104], Clinical BERT [105], DocBERT [109], RoBERT [110], ToBERT [110]
Attention Based
DiSan [93], ReSan [94], Bi-BloSAN [96], XLM [89]
Multimodal Embeddings

# BIBLIOGRAFÍA

[1] Almeida, F., & Xexéo, G. (2023). Word Embeddings: A Survey. https://arxiv.org/abs/1901.09069

[2] Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. Information Fusion, 89, 418–436. https://doi.org/https://doi.org/10.1016/j.inffus.2022.08.024