Redes Neuronales Artificiales en Procesamiento de Lenguaje Natural

Seminario de Aprendizaje Profundo aplicado al Lenguaje Natural

Abril, 2016

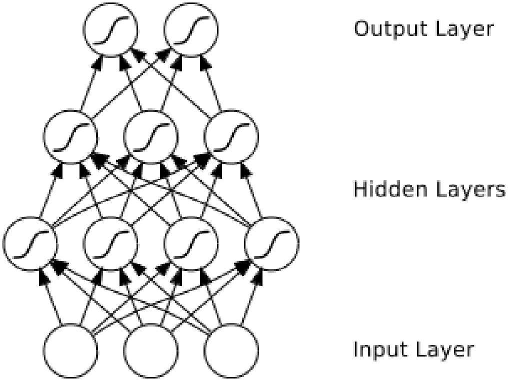
Contenido

1. Introducción
2. Representaciones Distribuidas en PLN
   - Palabras
   - Oraciones
   - Párrafos
3. Aplicaciones de RNA en PLN
   - Modelos de Lenguaje
   - Detección de Paráfrasis
   - Análisis Sintáctico
   - Análisis de Sentimiento
   - Traducción Automática
   - Reconocimiento del Habla
   - Question Answering

Motivación

- En las *features* manuales generalmente falta información. Cuesta diseñarlas y evaluarlas. Usan recursos externos.
- Es posible aprender representaciones distribuidas (*representation learning*) de la información. Representar palabras, imagenes, sonidos, etc. con vectores de dimensión alta.
- Las redes neuronales artificiales permiten crear representaciones distribuidas y utilizarlas para resolver tareas.

Red feed-forward



(Imagen tomada del libro "Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves.)

NLP (almost) from scratch. (Collober, 2011)

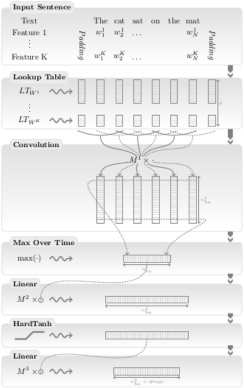


Figure 2: Sentence approach network.

(Imagen tomada del artículo "Natural Language Processing (almost) from Scratch "(Collobert et al., 2011))

NLP (almost) from scratch. (Collober, 2011)

- En todas las tareas propuestas se superó o se estuvo próximo al estado del arte.

Task		Benchmark	SENNA
Part of Speech (POS)	(Accuracy)	97.24 %	97.29 %
Chunking (CHUNK)	(F1)	94.29 %	94.32 %
Named Entity Recognition (NER)	(F1)	89.31 %	89.59 %
Parse Tree level 0 (PT0)	(F1)	91.94 %	92.25 %
Semantic Role Labelling (SRL)	(F1)	77.92 %	75.55 %

(Tabla tomada del artículo "Natural Language Processing (almost) from Scratch "(Collobert et al., 2011))

NLP (almost) from scratch. (Collober, 2011)

- No utiliza *features* manuales.
- No utiliza recursos lingüístcos (además de los conjuntos de entrenamiento).
- Se obtienen representaciones distribuidas de las palabras (tabla de *lookup*).
- Las mismas representaciones son utilizadas en todas las tareas.

Representaciones distribuidas de las palabras

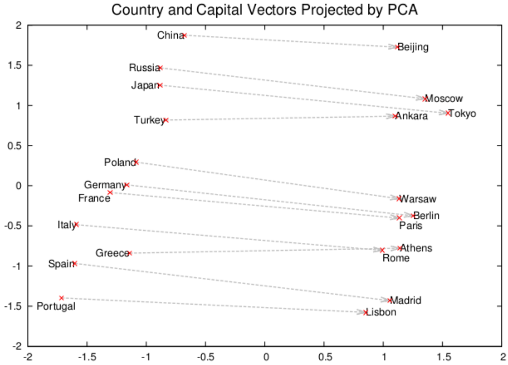
- Las representaciones distribuidas pueden interpretarse como puntos en un espacio vectorial de dimensión alta (Ej. 300).
- Las palabras relacionadas tienen representaciones cercanas.
- Algunas relaciones entre palabras preservan *offsets*

(Imagen tomada del artículo "Linguistic Regularities in Continuous Space Word Representations"(Mikolov et al., 2013))

"Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves "Natural Language Processing (almost) from Scratch"(Collobert et al., 2011) "Linguistic Regularities in Continuous Space Word Representations"(Mikolov et al., 2013)

Representaciones distribuidas de las palabras

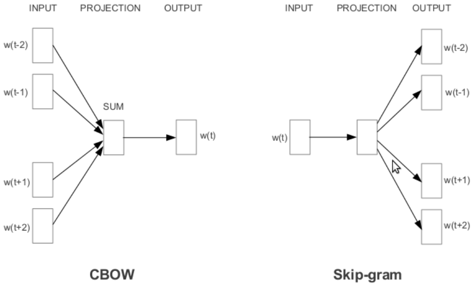
- También ocurre con relaciones semánticas.



(Imagen tomada del artículo "Distributed Representations of Words and Phrases and their Compositionality"(Mikolov et al., 2013))

Representar palabras con RNA

- Con modelos neuronales es posible construir buenas representaciones.
  - Ej. Skip-Gram y CBOW (Mikolov et al., 2013)



(Imagen tomada del artículo .Efficient Estimation of Word Representations in Vector Space "(Mikolov et al., 2013))

Modelos de conteo

- También se pueden obtener representaciones distribuidas contando. Por ejemplo, construyendo una matriz palabra-contexto y factorizándola con *SVD*. (*LSA* (Deerwester, 1990))
- Los representaciones basadas en redes neuronales parecían dar mejores resultados que las de conteo. (Baroni, 2014)

Modelos de conteo

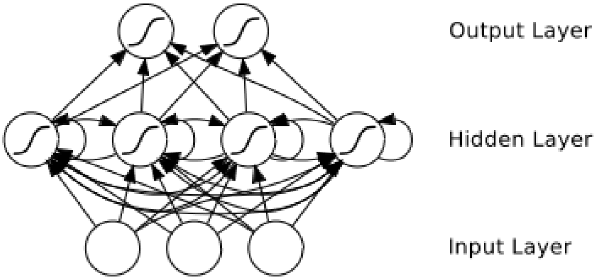
- Skip-gram (con *negative sampling*) converge a una factorización de la matrix palabra-contexto cuyas celdas son valores de *Pointwise Mutual Information* (PMI) desplazados por una constante global que depende de la cantidad de ejemplos negativos (Levy, 2014).

- Esto permitió mejorar los modelos de conteo obteniendo resultados comparables a los de predicción (Levy, 2015).

Representación de oraciones

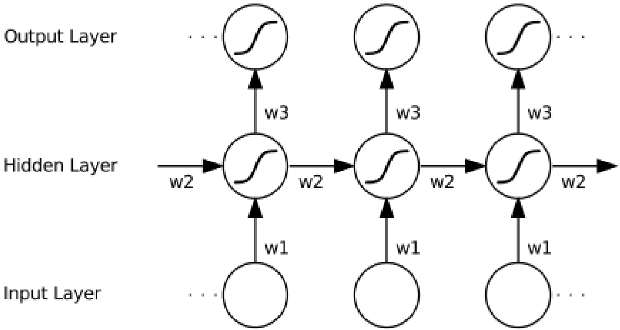
- Es posible representar el significado de una oración con un vector?
- Utilizar las representaciones de las palabras para construir un vector (de dimensión fija) que represente el significado de la oración.
- Diversas propuestas: redes convolutivas, redes recurrentes, autoencoders recursivos, etc.

Red Recurrente



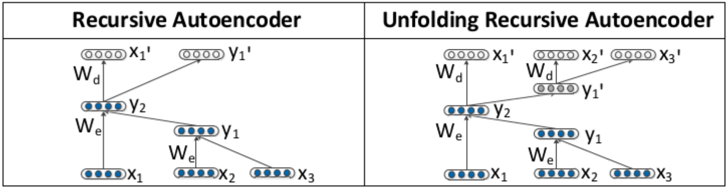
(Imagen tomada del libro "Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves.)

Red Recurrente (unfolded)



(Imagen tomada del libro "Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves.)

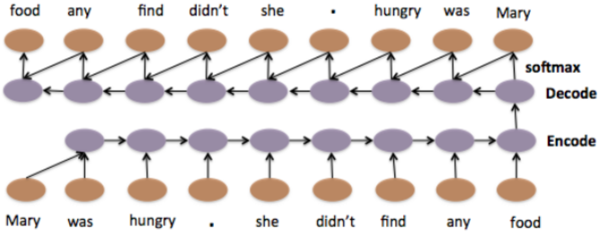
Recursive Autoencoder y Unfolding Recursive Autoencoder



(Imagen tomada del artículo "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection "(Socher et al., 2011))

Representación de párrafos

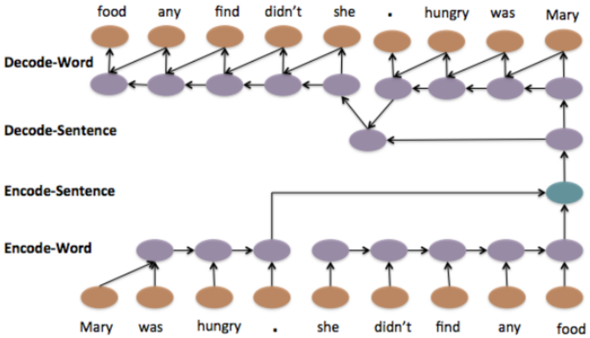
- Representación usando un autoencoder recursivo.



- (Imagen tomada del artículo .A Hierarchical Neural Autoencoder for Paragraphs and Documents "(Li et al., 2015))

Representación de párrafos

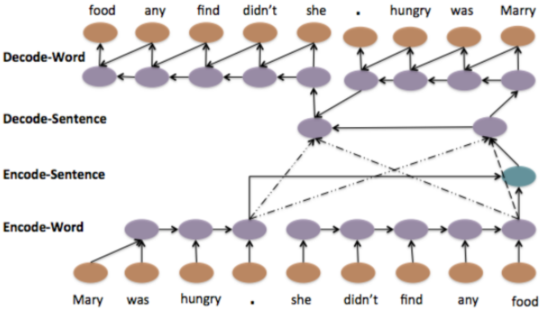
- Representación usando 2 niveles de *RAEs*.



- (Imagen tomada del artículo .A Hierarchical Neural Autoencoder for Paragraphs and Documents "(Li et al., 2015))

Representación de párrafos

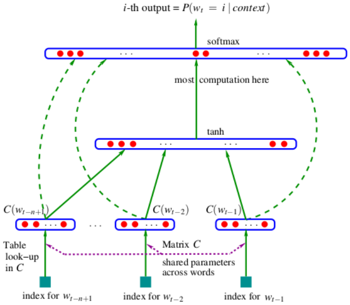
- Representación usando una jerarquía de autoencoders recurrentes (*with attention*).



- (Imagen tomada del artículo .A Hierarchical Neural Autoencoder for Paragraphs and Documents "(Li et al., 2015))

Modelos de Lenguaje con FFN

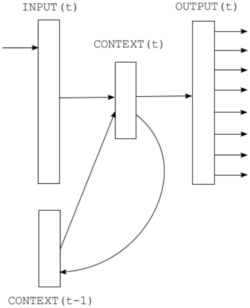
- Modelos de lenguaje con redes *feed forward* (Bengio et al., 2003).



- (Imagen tomada del artículo .A Neural Probabilistic Language Model"(Bengio et al., 2003))

Modelos de Lenguaje con RNN

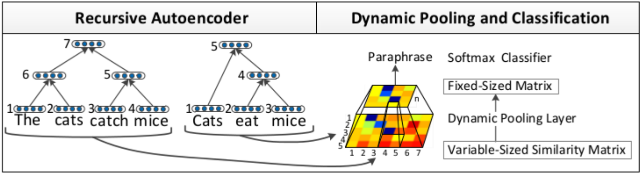
- Modelos de lenguaje con redes recurrentes (Mikolov et al., 2010).



- (Imagen tomada del artículo Recurrent neural network based language model"(Mikolov et al., 2010))

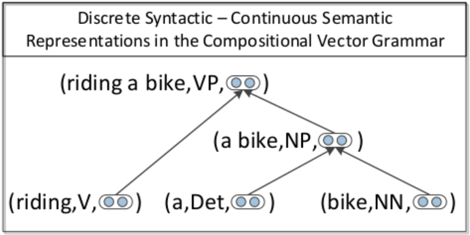
Detección de Paráfrasis

- Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection (Socher et al., 2011)



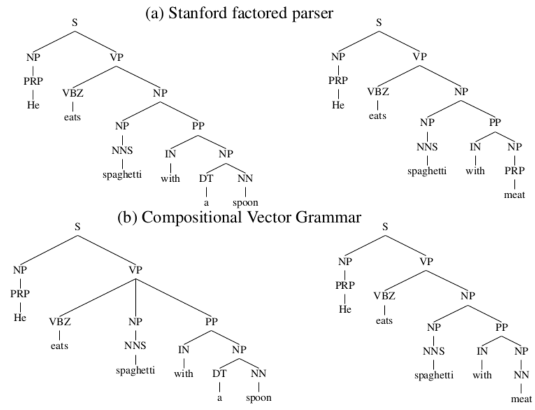
- (Imagen tomada del artículo "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection "(Socher et al., 2011))

Análisis Sintáctico

- Compositional Vector Grammars (Socher et al., 2013) Usa *word vectors* para considerar información semántica.
- 

(Imagen tomada del artículo "Parsing with Compositional Vector Grammars"(Socher et al., 2013))

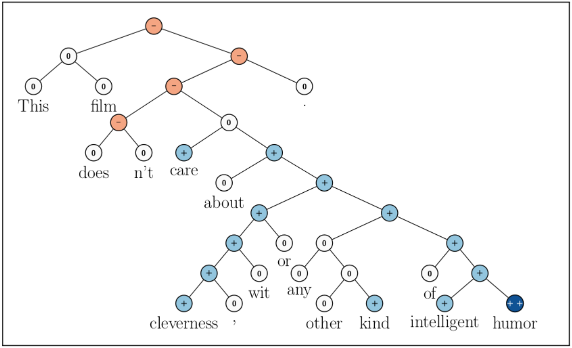
Análisis Sintáctico



(Imagen tomada del artículo "Parsing with Compositional Vector Grammars"(Socher et al., 2013))

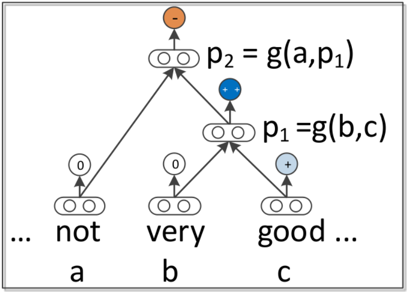
Análisis de Sentimiento

Vectores en nodos intermedios (Socher et al., 2013)



(Imagen tomada del artículo Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank"(Socher et al., 2013))

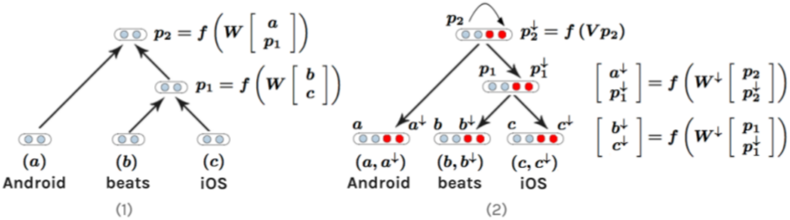
Análisis de Sentimiento



(Imagen tomada del artículo Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank"(Socher et al., 2013))

Análisis de Sentimiento

Global Belief Recursive Neural Networks (Paulus et al., 2014)



(Imagen tomada del artículo "Global Belief Recursive Neural Networks"(Paulus et al., 2014))

Análisis de Sentimiento

Classifier	Feature Sets	Twitter 2013 (F1)	SMS 2013 (F1)
SVM	stemming, word cluster, SentiWordNet score, negation	85.19	88.37
SVM	POS, lexicon, negations, emoticons, elongated words, scores, syntactic dependency, PMI	87.38	85.79
SVM	punctuation, word n-grams, emoticons, character n-grams, elongated words, upper case, stopwords, phrase length, negation, phrase position, large sentiment lexicons, microblogging features	88.93	88.00
GB-RNN	parser, unsupervised word vectors (ensemble)	**89.41**	**88.40**

Table 1: Comparison to the best Semeval 2013 Task 2 systems, their feature sets and F1 results on each dataset for predicting sentiment of phrases in context. The GB-RNN obtains state of the art performance on both datasets.

(Imagen tomada del artículo "Global Belief Recursive Neural Networks"(Paulus et al., 2014))

Análisis de Sentimiento

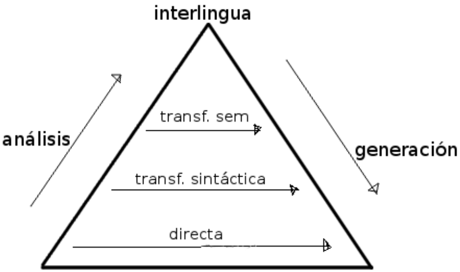
Method	Fine-grained	Binary
RAE (Socher et al., 2013)	43.2	82.4
MV-RNN (Socher et al., 2013)	44.4	82.9
RNTN (Socher et al., 2013)	45.7	85.4
DCNN (Blunsom et al., 2014)	48.5	86.8
Paragraph-Vec (Le and Mikolov, 2014)	48.7	87.8
CNN-non-static (Kim, 2014)	48.0	87.2
CNN-multichannel (Kim, 2014)	47.4	**88.1**
DRNN (Irsoy and Cardie, 2014)	49.8	86.6
LSTM	46.4 (1.1)	84.9 (0.6)
Bidirectional LSTM	49.1 (1.0)	87.5 (0.5)
2-layer LSTM	46.0 (1.3)	86.3 (0.6)
2-layer Bidirectional LSTM	48.5 (1.0)	87.2 (1.0)
Dependency Tree-LSTM	48.4 (0.4)	85.7 (0.4)
Constituency Tree-LSTM		
– randomly initialized vectors	43.9 (0.6)	82.0 (0.5)
– Glove vectors, fixed	49.7 (0.4)	87.5 (0.8)
– Glove vectors, tuned	**51.0** (0.5)	88.0 (0.3)

Table 2: Test set accuracies on the Stanford Sentiment Treebank. For our experiments, we report mean accuracies over 5 runs (standard deviations in parentheses). **Fine-grained:** 5-class sentiment classification. **Binary:** positive/negative sentiment classification.

(Tabla tomada del artículo Ïmproved Semantic Representations From Tree-Structured Long Short-Term Memory Networks"(Tai et al., 2015))

Traducción Automática

- Triángulo de Vauquois



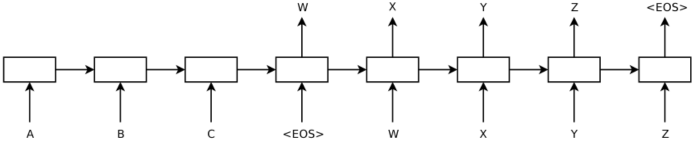
-

- interlingua = representaciones distribuidas ?

Traducción Automática

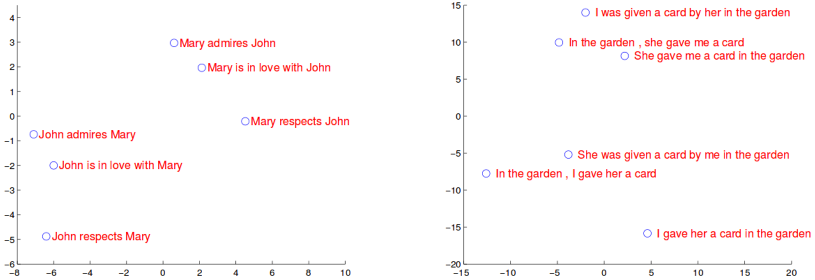
- Se utiliza una red para obtener una representación de la oración. (*encoder*)
- A partir de la representación se decodifica la oración en el lenguaje destino. (*decoder*)
- LSTMs (Sutskever et al. 2014), R2NN (Liu et al., 2014), Redes Convolutivas (Cho et al., 2014), ...

Traducción Automática



(Imagen tomada del artículo "Sequence to Sequence Learning with Neural Networks"(Sutskever et al. 2014))

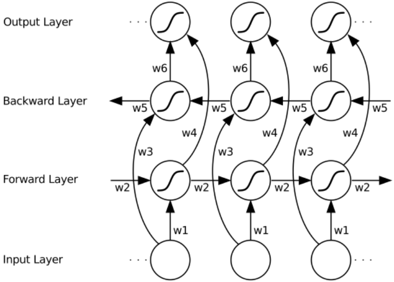
Traducción Automática



(Imagen tomada del artículo "Sequence to Sequence Learning with Neural Networks"(Sutskever et al. 2014))

Reconocimiento del Habla

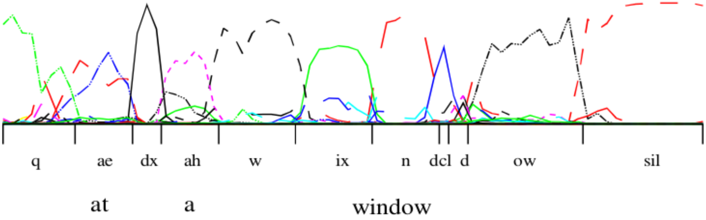
- Se han utilizado redes recurrentes bidireccionales (Graves et al., 2005).



- (Imagen tomada del libro "Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves.)

Reconocimiento del Habla

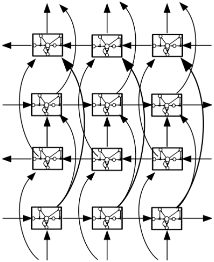
- Las redes bidireccionales usan contexto pasado y futuro para realizar predicciones.



- (Imagen tomada del libro "Supervised Sequence Labelling with Recurrent Neural Networks"de Alex Graves.)

Reconocimiento del Habla

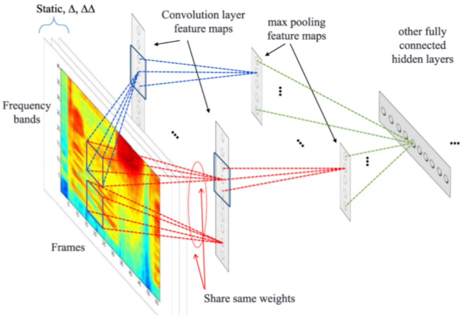
- Se han utilizado redes recurrentes bidireccionales profundas (Graves et al., 2013).



- (Imagen tomada del libro "Hybrid Speech Recognition with Deep Bidirectional LSTM"(Graves et al., 2013))

Reconocimiento del Habla

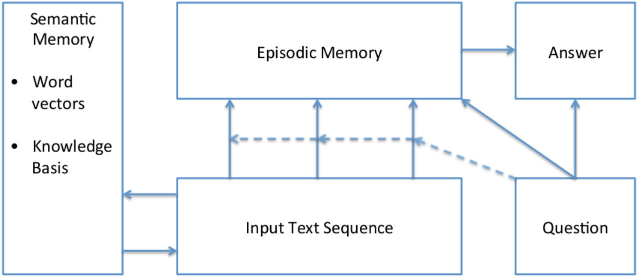
- También redes convolutivas (Abdel-Hamid et al., 2014).



- (Imagen tomada de Çonvolutional Neural Networks for Speech Recognition"(Abdel-Hamid et al., 2014))

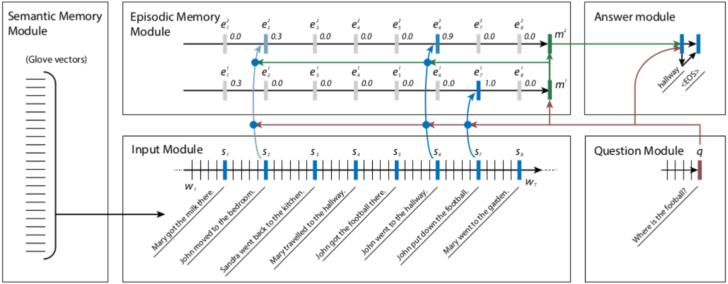
Question Answering

- Dynamic Memory Networks para PLN (Kumar et al., 2015).



- (Imagen tomada de .Ask Me Anything: Dynamic Memory Networks for Natural Language Processing"(Kumar et al., 2015))

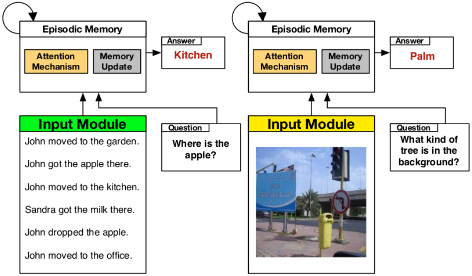
Question Answering



(Imagen tomada de .Ask Me Anything: Dynamic Memory Networks for Natural Language Processing"(Kumar et al., 2015))

Question Answering

- Dynamic Memory Networks para responder en textos e imágenes (Kumar et al., 2015).



(a) Text Question-Answering (b) Visual Question-Answering

- (Imagen tomada de "Dynamic Memory Networks for Visual and Textual Question Answering"(Xiong et al., 2016))

Preguntas?

- Preguntas?

Referencias I

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014).
Convolutional neural networks for speech recognition.
IEEE/ACM Trans. Audio, Speech & Language Processing, 22(10):1533–1545.

Baroni, M., Dinu, G., and Kruszewski, G. (2014).
Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 238–247.

Referencias II

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
Journal of Machine Learning Research, 3:1137–1155.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011).
Natural language processing (almost) from scratch.
Journal of Machine Learning Research, 12:2493–2537.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41.

Referencias III

Graves, A. (2012).
Supervised Sequence Labelling with Recurrent Neural Networks, volume 385 of *Studies in Computational Intelligence*.
Springer.

Graves, A., Jaitly, N., and Mohamed, A. (2013).
Hybrid speech recognition with deep bidirectional LSTM.
In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 273–278.

Referencias IV

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Gulrajani, I., and Socher, R. (2015).
Ask me anything: Dynamic memory networks for natural language processing.
CoRR, abs/1506.07285.

Levy, O. and Goldberg, Y. (2014).
Neural word embedding as implicit matrix factorization.
In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185.

Referencias V

Levy, O., Goldberg, Y., and Dagan, I. (2015).
Improving distributional similarity with lessons learned from word embeddings.
TACL, 3:211–225.

Li, J., Luong, M., and Jurafsky, D. (2015).
A hierarchical neural autoencoder for paragraphs and documents.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1106–1115.

Referencias VI

Liu, S., Yang, N., Li, M., and Zhou, M. (2014).
A recursive recurrent neural network for statistical machine translation.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1491–1500.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient estimation of word representations in vector space.
In Proceedings of Workshop at ICLR.

Referencias VII

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010).
Recurrent neural network based language model.
In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
In Proceedings of NIPS.

Referencias VIII

Mikolov, T., tau Yih, W., and Zweig, G. (2013c).
Linguistic regularities in continuous space word representations.
In Proceedings of NAACL HLT.

Paulus, R., Socher, R., and Manning, C. D. (2014).
Global belief recursive neural networks.
In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2888–2896.

Referencias IX

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013a).
Parsing with compositional vector grammars.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 455–465.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011).
Dynamic pooling and unfolding recursive autoencoders for paraphrase detection.
In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 801–809.

Referencias X

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b).
Recursive deep models for semantic compositionality over a sentiment treebank.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.
In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Referencias XI

Tai, K. S., Socher, R., and Manning, C. D. (2015).
Improved semantic representations from tree-structured long short-term memory networks.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566.

Xiong, C., Merity, S., and Socher, R. (2016).
Dynamic memory networks for visual and textual question answering.
CoRR, abs/1603.01417.