Моделирование распознавания семантических ролей для русского языка

Екатерина Гавришина

Руководитель: О. Н. Ляшевская

8 июня, 2021

Москва, НИУ ВШЭ

Содержание

SRL

— распознавание семантических ролей (semantic role labeling), или неглубокий семантический парсинг.

Семантические участники ситуации, или "глубинные падежи" (Fillmore 1968).

“ *Райт* *владеет* *чешским языком.*

Агенс Предикат Тема

*Пример из FrameBank (Кашкин, Ляшевская 2013)

Задачи

Цель: реализовать последовательную автоматическую разметку семантических ролей на материале СинТагРус (Дяченко и др. 2015).

1) предобработка обучающего корпуса
2) разработка метрик оценки качества модели
3) построение нейросетевых моделей
4) сравнение качества работы моделей
5) распознавание семантических ролей на тестовом корпусе
6) анализ результатов

Обзор ресурсов

FrameNet (Baker et al. 1998)

Фрейм “commerce_buy”: *купить, покупатель, клиент, покупка, приобрести, приобретение.*

Роли: Покупатель, Товар

Продавец, Деньги и т.д.

PropBank (Palmer et al. 2005)

Глагол “buy”

Роли: A-0 “buyer”, A-1 “thing bought”, A-2 “seller”, A-3 “price paid” и т.д.

Обзор ресурсов

FrameBank (Кашкин, Ляшевская 2013)

— основан на конструкциях конкретных лексем, а не на типовых ситуациях (фреймах)

Russian PropBank (Moeller et al. 2020)

— автоматическое проецирование английских семантических ролей на русские тексты

Обзор методов

Основные подходы:

- правиловые методы (Sokirko 2001; Shelmanov, Smirnov 2014)
- статистические методы (Gildea, Jurafsky 2002)
- нейронные сети (Collobert et al. 2011; Fonseca, Rosa 2013; Foland, Martin 2015; Shelmanov, Devyatkin 2017; Li et al. 2020)
- механизм внимания (Strubell et al. 2018)

Данные для обучения

FrameBank + SyntaxNet (Shelmanov, Devyatkin 2017)

— морфологическая, семантическая и автоматически добавленная синтаксическая разметка

- формат JSON
- **7952** предложения с семантической разметкой
- **10%** токенов имеют семантическую роль
- **34%** присвоена роль предиката, **7%** — роль агенса, **6%** — роль пациенса, **5%** — роль темы

Данные для тестирования

СинТагРус (Дяченко и др. 2015)

— эталонная синтаксическая разметка в формате Universal Dependencies (UD)

- формат CoNLL-U
- **6491** предложение (тестовый подкорпус)

Предобработка

Добавлены признаки *lemma_parent*, *pos_parent*, *parent*.

FullIndex	index	lemma	POS	feat_p	link_name	lemma_parent	pos_parent	parent	Role
36709_4	1	но	CCONJ	_	cc	недосуг	VERB	4	_
36709_4	2	хозяин	NOUN	Animacy=Anim|Case=Dat|Gender=Masc|Number=Sing	nmod	недосуг	VERB	3	субъект социального отношения
36709_4	3	нынче	ADV	Degree=Pos	advmod	недосуг	VERB	2	_
36709_4	4	быть	AUX	Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Te...	cop	недосуг	VERB	1	_
36709_4	5	недосуг	VERB	Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing	root	_	_	0	_
36709_4	6	с	ADP	_	case	он	PRON	1	_
36709_4	7	он	PRON	_	dobj	баловать	VERB	1	пациенс социального отношения
36709_4	8	баловать	VERB	Aspect=Imp|VerbForm=Inf	xcomp	недосуг	VERB	-3	предикат
36709_4	9	,	PUNCT	_	punct	баловать	VERB	-1	_
36709_4	10	он	PRON	_	nsubj	спешить	VERB	3	_
36709_4	11	в	ADP	_	case	дом	NOUN	1	_
36709_4	12	дом	NOUN	Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing	dobj	спешить	VERB	1	_
36709_4	13	спешить	VERB	Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Te...	conj	недосуг	VERB	-8	_
36709_4	14	.	PUNCT	_	punct	недосуг	VERB	-9	_

Рис. 1. Формат представления данных.

Обучающая (**60%**, **4770** предложений), валидационная (20%) и тестовая (20%) выборки.

Эмбеддинги

1) модель, обученная на НКРЯ и Википедии (788 млн слов, размерность вектора 300)

2) модель, обученная на русскоязычных новостях (2,6 млрд слов, размерность вектора 300)

3) мультилингвальная модель Google (16 языков, размерность вектора 512) (Yang et al. 2019)

Модели

Слои: эмбеддинги, BiLSTM, полносвязные (Dense)

Функция потерь: категориальная кросс-энтропия

Оптимизатор: алгоритм Adam

- Отдельные входы для категориальных признаков (Shelmanov, Devyatkin 2017)
- Использование информации о вершине
- Dropout
- 4 модели

Архитектура

1) Модель с четырьмя входами без учёта информации о вершине и одним Dense-слоем (рис. 2)

2) Модель с входами для семи признаков и одним Dense-слоем

3) Модель с добавлением Dense-слоя на каждом входе до конкатенации (рис. 3)

4) Модель с добавлением ещё одного Dense-слоя после BiLSTM перед выходным слоем

Архитектура

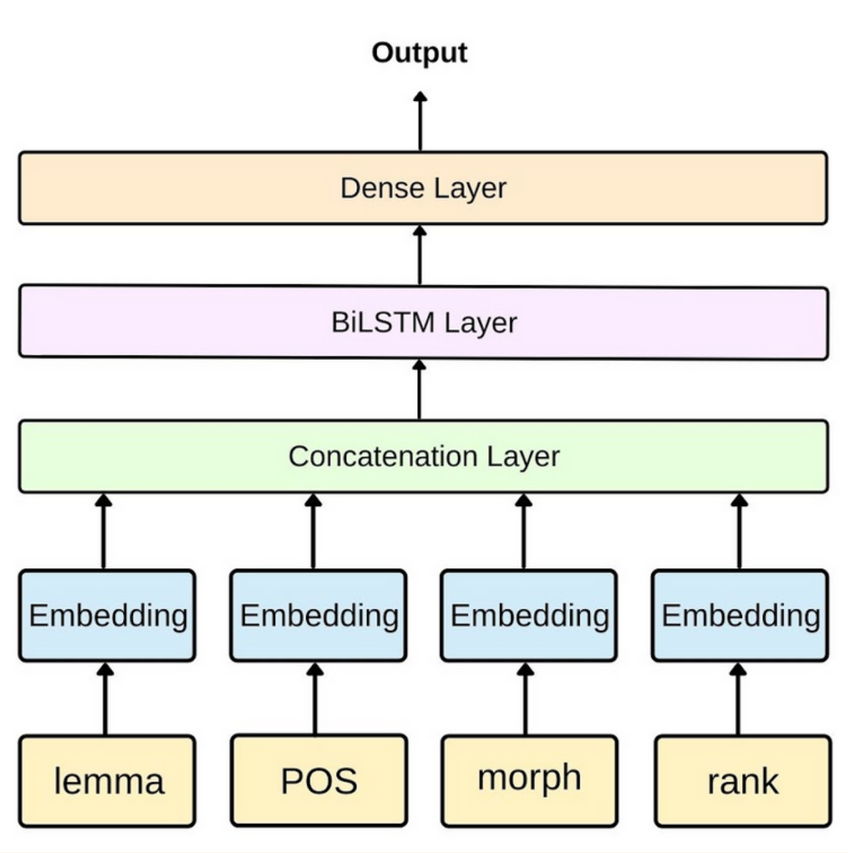

Рис. 2. Архитектура простейшей модели.

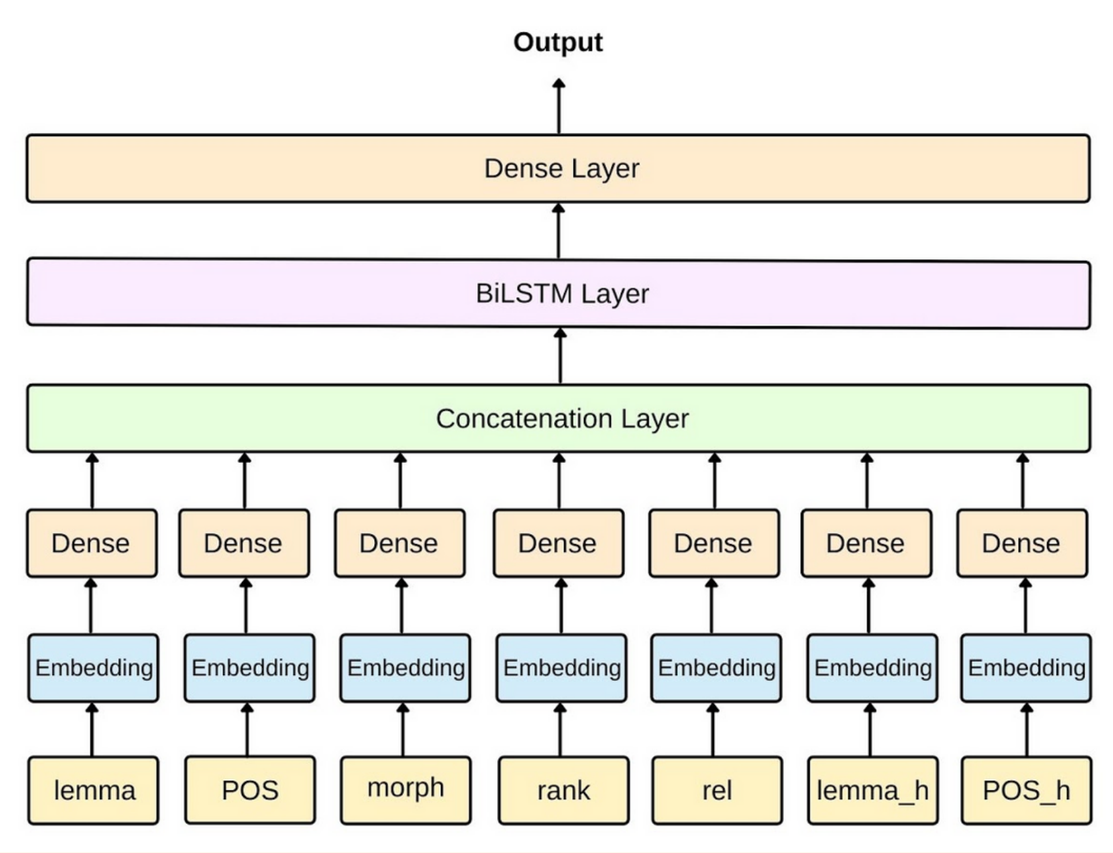

Рис. 3. Архитектура модели с полносвязным слоем до конкатенации.

Метрики

Accuracy, Precision (точность), Recall (полнота)

Разработанные метрики:

- **Labeled Accuracy:** позиции, которым модель присвоила класс, отличный от “_” и от “pad”
- точность и полнота для роли **предиката, агенса** и **пациенса** (только на тестовых данных)

Сравнение моделей

Модели	Accuracy	Precision	Recall	Labeled Accuracy
Модель 1	0.9054	0.9123	0.8997	0.6936
Модель 2	0.9088	0.9252	0.8947	0.5997
Модель 3	0.9172	0.9395	0.8990	0.6588
Модель 4	0.9156	0.9409	0.9008	0.6812

Таблица 1. Значения основных метрик на тестовой выборке для моделей.

Модели	Предикат		Агенс		Пациенс	
	Precision	Recall	Precision	Recall	Precision	Recall
Модель 1	0.7063	0.3221	0.25	0.0035	0.0	0.0
Модель 2	0.6351	0.4968	0.4884	0.0734	0.4839	0.06
Модель 3	0.7265	0.6646	0.4516	0.1958	0.4133	0.124
Модель 4	0.7512	0.6744	0.3529	0.0839	0.2475	0.100

Таблица 2. Точность и полнота частотных классов для моделей.

Результаты

6941 предложение (117327 токенов) —> **1879** ролей

1. Модель в большей мере подстраивается под лексику

(1) *Свое лицо он знал наизусть и никогда не* ***пугал*****[предикат]** *остатками грима* ***кондукторов*****[причина]** *последних ночных трамваев.*

Результаты

6941 предложение (117327 токенов) —> **1879** ролей

1. Модель в большей мере подстраивается под лексику

2. Высокое качество работы модели в отношении ролей с ярко выраженными лексическими особенностями

(2) *Именно ему* ***я*****[говорящий]** ***задал*****[предикат]** *главный интересующий меня вопрос.*

(3) *Потенциальных многодетных* ***родителей*****[субъект психологического состояния]** ***смущает*****[предикат]** ***то*****[причина]**, *что деньги нельзя потратить на лечение ребенка или покупку садового участка.*

Результаты

6941 предложение (117327 токенов) —> **1879** ролей

1. Модель в большей мере подстраивается под лексику
2. Высокое качество работы модели в отношении ролей с ярко выраженными лексическими особенностями
3. Если размечен пациенс, то размечен и агенс, если он есть

(4) *Схватив работающие пилы,* ***спортсмены*****[агенс]** *начинают очень быстро* ***отрезать*****[предикат]** *у* ***бревен*****[пациенс]** *ровные диски.*

(5) *От холода* ***слона*****[пациенс]** *можно* ***спасти*****[предикат]**, *если дать ему напиться густого красного вина.*

Результаты

6941 предложение (117327 токенов) —> **1879** ролей

1. Модель в большей мере подстраивается под лексику
2. Высокое качество работы модели в отношении ролей с ярко выраженными лексическими особенностями
3. Если размечен пациенс, то размечен и агенс, если он есть
4. Роль темы при фазовых (и подобных) глаголах

(6) *Загородный дом можно сдать в аренду — и у вас* ***появляется*****[предикат]** *весьма приличный* ***доход*****[тема]**.

(7) *А тем времени вокруг проекта* ***продолжались*****[предикат]** *научные и околонаучные* ***споры*****[тема]**.

Заключение

1) Сопоставлены и проанализированы различные архитектуры многослойной сети

2) Подтверждена гипотеза о важности синтаксической информации

3) Использована идея разделения входного слоя

4) Разработаны методы оценки модели

Дальнейшее развитие:

- применение механизма внимания (или “само-внимания”)
- подбор различных комбинаций признаков и параметров модели
- специализированная аугментация данных

Код и материалы: https://github.com/kategavrishina/srl-modeling

Список литературы

Гавришина 2020 — Гавришина Е.И. Эталонный корпус русского языка для универсального семантического парсинга. Курсовая работа, НИУ ВШЭ. 2020.

Дяченко и др. 2015 — Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) //*Сборник «Национальный корпус русского языка: 10 лет проекту». Труды Института русского языка им. В.В. Виноградова.* Москва, 2015. Вып. 6. С. 272–299.

Кашкин, Ляшевская 2013 — Кашкин Е. В., Ляшевская О. Н. Семантические роли и сеть конструкций в системе FrameBank // *Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции Диалог.* 2013. С. 297-311.

Baker et. al 1998 — Baker C. F., Fillmore C. J., Lowe J. B. The berkeley framenet project //*Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics.* 1998. P. 86–90.

Cai et al. 2018 — Cai, J., He, S., Li, Z., Zhao, H. A Full End-to-End Semantic Role Labeler, Syntax-agnostic Over Syntax-aware? //*arXiv preprint arXiv:1808.03815.* 2018.

Collobert et al. 2011 — Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch //*Journal of machine learning research.* V. 12, №. Aug, 2011. P. 2493–2537.

Ezen-Can 2020 — Ezen-Can A. A Comparison of LSTM and BERT for Small Corpus //*arXiv preprint arXiv:2009.05451*. 2020.

Fillmore 1968 — Fillmore C. J. The case for case //*Universals in Linguistic Theory / Ed. by Emmon Bach, Robert T. Harms.* 1968. P. 1–88.

Foland, Martin 2015 — Foland, W., Martin, J. H. Dependency-based semantic role labeling using convolutional neural networks //*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics.* 2015. P. 279-288.

Fonseca, Rosa 2013 — Fonseca, E. R., Rosa, J. L. G. A two-step convolutional neural network approach for semantic role labeling //*The 2013 International Joint Conference on Neural Networks (IJCNN).* 2013. P. 1-7. IEEE.

Gildea, Jurafsky 2002 — Gildea D., Jurafsky D. Automatic labeling of semantic roles //*Computational linguistics.* V. 28, №. 3, 2002. P. 245–288.

Marcheggiani, Titov 2017 — Marcheggiani, D., Titov, I. Encoding sentences with graph convolutional networks for semantic role labeling. //*arXiv preprint arXiv:1703.04826.* 2017.

Moeller et al. 2020 — Moeller, S., Wagner, I., Palmer, M., Conger, K., Myers, S. The Russian PropBank //*Proceedings of The 12th Language Resources and Evaluation Conference.* 2020. P. 5995– 6002.

Palmer et al. 2005 — Palmer M., Gildea D., Kingsbury P. The proposition bank: An annotated corpus of semantic roles //*Computational linguistics.* V. 31, №. 1, 2005. P. 71–106.

Li et al. 2020 — Li, Z., Zhao, H., He, S., Cai, J. Syntax Role for Neural Semantic Role Labeling //*arXiv preprint arXiv:2009.05737*. 2020.

Shelmanov, Devyatkin 2017 — Shelmanov A. O., Devyatkin D. A. Semantic role labeling with neural networks for texts in Russian //*Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" 2017*. V. 16, 2017. P. 245–256.

Shelmanov, Smirnov 2014 — Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts //*Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog*. V. 13, №. 20, 2014. P. 607–620.

Sokirko 2001 — Sokirko, A. A short description of Dialing Project. *Retrieved from http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html.* 2001.

Strubell et al. 2018 — Strubell, E., Verga, P., Andor, D., Weiss, D., McCallum, A. Linguistically-informed self-attention for semantic role labeling //*arXiv preprint arXiv:1804.08199.* 2018.

Yang et al. 2019 — Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Kurzweil, R. Multilingual universal sentence encoder for semantic retrieval //*arXiv preprint arXiv:1907.04307*. 2019.

Zhou, Xu 2015 — Zhou J., Xu W. End-to-end learning of semantic role labeling using recurrent neural networks //*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.* 2015. P. 1127-1137.

Спасибо за внимание!

BiLSTM

SRL: (Zhou, Xu 2015; Marcheggiani, Titov 2017; Cai et al. 2018; Li et al. 2020)

BiLSTM vs Bert: (Ezen-Can 2020)

- учитывает контекст
- работает в "двух направлениях"
- используется в seq2seq моделях