электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.934.1

Устранение лексической неоднозначности при решении задачи машинного статического перевода

***Маланин Г.П.**, студент*
Россия, 1050005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»

Научный руководитель: Барышникова М.Ю., к.п.н, доцент
Россия, 1050005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»
irudakov@bmstu.ru

Введение

Задача устранения лексической неоднозначности является одной из центральных задач обработки текста и компьютерной лингвистики в целом.

Под устранением лексической неоднозначности будем понимать определение смысла конкретного слова с учетом контекста, который его окружает.

Если не учитывать пунктуацию, можно представить текст T как последовательность слов $(w_1 w_2, .., w_n)$. Тогда задача устранения лексической неоднозначности может быть формально определена как назначение соответствующего смысла sense(s) для всех слов из последовательности T.

Применительно к задаче машинного перевода устранение лексической неоднозначности можно определить, как выбор наиболее подходящего варианта перевода слова с исходного языка на целевой

Существующие подходы

Для устранения многозначности в машинном переводе в основном применяется подход обучения с учителем.

На вход системы подается размеченный корпус, состоящий из текста на исходном языке, в котором каждое неоднозначное слово помечено переводом на целевой язык. В этом случае задача устранения неоднозначности сводится к стандартной классификации, где классы — это возможные значения (перевод) слова. На вход такому классификатору подается локальный и/или глобальный контекст (контекст в данном случае — это не только окружающие слова, но и возможно их части речи).

Также возможны подходы, основанные на частичном обучении с учителем. При этом подразумевается, что существует небольшой размеченный корпус и большой неразмеченный. В этом случае также используются классификатор, который сначала обучается на размеченном корпусе, а потом из неразмеченного корпуса выбирает слова, чьи классы (переводы) установлены с определенным уровнем «точности» и добавляет в размеченный корпус, так продолжается пока для всех слов из неразмеченного корпуса не будут установлены их значения.

Для подходов обучения с учителем и частичного обучения с учителем в основном применяется следующие алгоритмы классификации: наивный байесовский классификатор [1], *EM*–алгоритм (*expectation–maximization*) [2] и нейронные сети [3].

Главным недостатком таких подходов является трудоемкость разметки корпуса и малая точность классификатора при недостаточном размере корпуса.

Учитывая проблемы существующих подходов, была предпринята попытка разработки метода, устраняющего лексическую неоднозначность на стороне целевого языка и использующего подход обучения без учителя.

Описание метода

Главным элементом разработанного метода является распределенное представление слов.

Распределенное представление слов — это вектор действительных чисел, используемый для различных задач обработки естественного языка. К преимуществам данного представления можно отнести компактность вектора и близость векторов слов, встречающихся в схожем контексте. То есть если слова «банк» и «финансы» расположены в тексте рядом друг с другом, то они будут иметь схожий вектор.

Впервые использование распределенного представления было представлено в 1986 году Румельхартом, Хинтоном и Вильямсом [4]. Успешно применена Бенгио в [5]. Существенно усовершенствована Милковым в [6].

В своей работе Милков вводит так называемую *CBOW* модель нейронной сети (рисунок 1). Отличительной ее чертой является большая скорость (100 миллиардов слов в день) обработки текста, а также большую точность по сравнению с методами *LSA* (латентно–семантический анализ) и *LDA* (латентное размещение Дирихле) [6].

В данной работе была использована *CBOW* модель нейронной сети. Для обучения сети использовался дамп русскоязычной Википедии. Википедия была выбрана по нескольким причинам: количество статей превышает 3 миллиона, каждая статья имеют

четкую тематику, лицензия Википедии позволяет использовать материалы для научных целей.

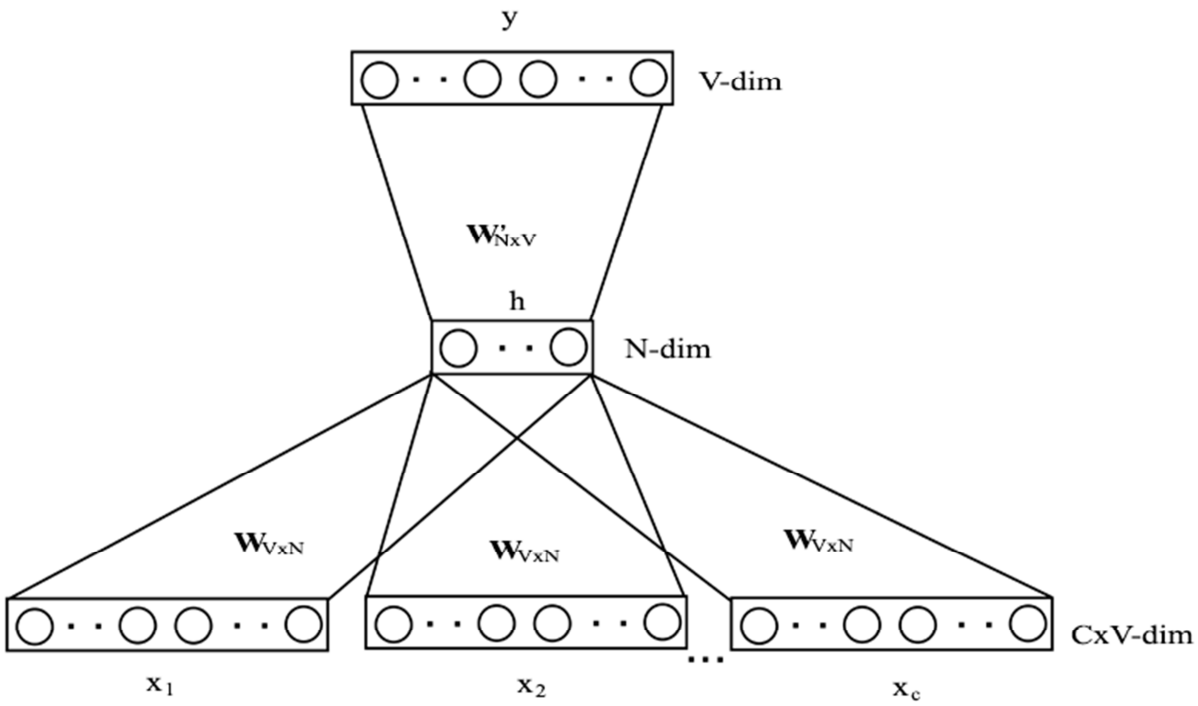

Рис. 1. *CBOW* нейронная сеть: **X** — входы (контекст); **C** — размер контекста; **V** размер словаря; N — размер вектора распределенного представления; **W, W'** — веса нейронной сети; **Y** — целевое слово, для которого определяется контекст

Для работы с нейронной сетью тексты Википедии были специально подготовлены: были убраны все знаки препинания, каждому уникальному слову был назначен свой идентификатор, представленный в виде двоичного числа (вектора) в котором только одна цифра равна единице, а все остальные равны нулю. Умножения этого вектора на веса нейронной сети ($W * x$), располагаемые между входным и скрытым слоем, и дают распределенное представление слова.

На вход нейронной сети может подаваться как локальный контекст, так и глобальный. Локальный контекст можно представить, как несколько соседних слов окружающих целевое.

Для выбора слов, которые отражают глобальный контекст (тематику) используется алгоритм tf–idf, при этом для более точных результатов перед использованием указанного алгоритма производится лемитизация всех слов. Под лемитизацией понимается приведение слова к его начальной форме (для существительного, например, это единственное число, именительный падеж).

На выход нейронной сети, в зависимости от того локальный или глобальный контекст используется, подается либо целевое слово, которое окружает контекст, либо все слова данного текста для которого определялся глобальный контекст.

Для обучения используются методы градиентного спуска.

Кроме распределенного представления слов целевого языка для работы метода необходим словарь переводов с исходного языка на целевой.

В итоге задача устранения неоднозначности сводится к выбору наилучших переводов для всех слов исходного текста, то есть слов, у которых распределенное представление максимально близко друг к другу.

Более формально решение задачи устранения неоднозначности может быть описано следующим образом: пусть исходный текст представлен упорядоченным набором слов $(s_1, .., s_n)$ где $\forall i\ s_i \in S$. Пусть S — множество слов исходного языка, а n количество слов в тексте. Каждое слово может иметь один или более переводов $Translate(s_i) \subseteq T$, где T множество слов целевого языка. Тогда задача сводится к поиску вектора $\hat{X} = (\hat{t}_1, .., \hat{t}_n)$ где $\forall i\ t_i \in Translate(s_i)$ при том, что:

$$\hat{X} = \underset{X}{\operatorname{argmin}}\, \sigma(X)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Distance^2(t_i, t_{centroid})}$$

$$Distance(t, t_{centroid}) = 2 - \left(1 + \frac{< t, t_{centroid} >}{|t||t_{centroid}|}\right)$$

$$t_{centroid} = \sum_{i=1}^{n} t_i$$

Distance — измененная косинусная мера, принимающая значения от 0 до 2.

Т.е. в такой постановке задача сводится к поиску векторов, имеющих наименьшее квадратичное отклонение, что позволяет заменить скалярные значения векторными величинами, вместо разности использовать измененную косинусную меру, принимающую значения в интервале от 0 до 2, а вместо среднего — сумму всех векторов (центроид) (рисунок 2).

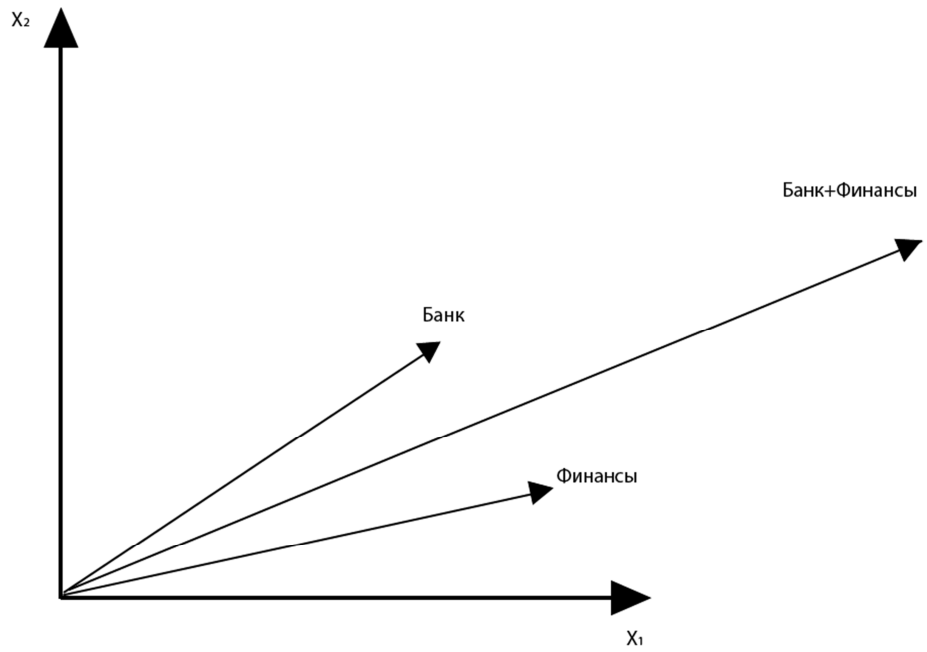

Рис. 2. Пример распределенного представления в виде векторов и их центроида

Так как в одном контексте могут встретиться несколько неоднозначных слов, данная задача может быть отнесена к задачам минимизации. Для ее решения был использован алгоритм имитации отжига.

Суть алгоритма имитации отжига заключается в итерационном поиске минимума заданной функции. На каждом шаге случайным образом выбираются значения аргументов функции, и если эти аргументы дают лучший результат минимизации функции по сравнению с текущим минимумом, то они принимаются за новый минимум, если нет, то выбор аргументов в качестве минимума осуществляется с определенной вероятностью, причем с каждым шагом применения алгоритма данная вероятность будет уменьшаться. Таким образом, на первых шагах выбор аргументов производится почти случайным образом, что позволяет исследовать функцию практически на всей области ее определения, однако со временем вероятность перехода в менее оптимальное состояние уменьшается, и алгоритм становится схож с методом градиентного спуска, обеспечивая постепенное достижение локального минимума.

Вероятность выбора новых аргументов X для функции F вычисляется по следующей формуле:

$$P(X_{i+1}|X_i) = \begin{cases} 1, F(X_{i+1}) - F(X_i) < 0 \\ e^{-\frac{F(X_{i+1}) - F(X_i)}{T(i+1)}}, F(X_{i+1}) - F(X_i) \geq 0 \end{cases},$$

$$\text{где } \forall i \, (T(i) > 0 \, \wedge \, T(i+1) \leq T(i))$$

Тестирование

Разработанный метод был протестирован на примере из ста фраз, для каждой из которых были представлены несколько вариантов перевода (правильный и неправильные). Тестирование показало, что при его использовании достигается точность перевода порядка 86 %. Примеры входных данных и результатов тестирования приведены в таблице.

Слово	**Перевод**	**Фраза**	**Среднеквадратичное отклонение распределенного представления слов**
bank	берег	Подплыл к берегу	0.302
	банк	Подплыл к банку	0.319
hood	капот	Капот автомобиля	0.195
	капюшон	Капюшон автомобиля	0.274
develop	разрабатывать	Разрабатывать участок земли	0.334
	конструировать	Конструировать участок земли	0.35

Заключение

В данной работе был представлен метод устранения лексической неоднозначности при решении задачи машинного статического перевода, основанный на распределенном представление слов и алгоритме имитации отжига. Разработанный метод является методом обучения без учителя. Данный подход не требует разметки обучающих данных, и обладает высокой точностью.

Список литературы

1. Yarowsky D., Florian R. Evaluating sense disambiguation across diverse parameter spaces. Available at: http://www.coli.uni-saarland.de/~kowalski/senseval2/yarowsky.pdf, accessed 21.04.2015.

2. Klein D.E., Bleda M.J., Manning D.C. Conditional structure versus conditional estimation in NLP models. Available at: http://nlp.stanford.edu/pubs/objective-functions.pdf, accessed 21.04.2015.
3. Veronis J., Ide N.M. Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. Available at: http://www.aclweb.org/anthology/C90-2067, accessed 21.04.2015.
4. Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by backpropagating errors. Available at: http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf, accessed 21.04.2015.
5. Bengio Y., Ducharme R., Pascal V., Jauvin C. A neural probabilistic language model. Available at: http://jmlr.csail.mit.edu/papers/volume3/bengio03a/bengio03a.pdf, accessed 21.04.2015.
6. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. Available at: http://arxiv.org/pdf/1301.3781.pdf, accessed 21.04.2015.