Применение лексических цепочек для разрешения лексической многозначности на основе Русского Викисловаря

Andrew Krizhanovsky

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences

Stanislav Tkach

Petrozavodsk State University

February 24, 2018

Человеческому языку присуща неоднозначность. В частности лексическая неоднозначность представлена во всех естественных языках. Например, в английском языке существительное «*plant*» может означать «*зеленое растение*» или «*завод*», аналогично французское слово «*feuille*» может иметь значение «*лист (растения или дерева)*» или «*лист бумаги*». Точное толкование многозначного слова может быть выбрано на основе контекста, в котором оно употребляется и соответственно задача выбора верного значения слова определяется как задача автоматического назначения наиболее подходящего для пользователя толкования данного слова в пределах контекста [5].

Лексическая многозначность (полисемия) — eto fundamiental'noie svoistvo iestiestviennykh iazykov: kazhdoie slovo mozhiet imiet' bolieie odnogho znachieniia.

Razrieshieniie lieksichieskoi mnoghoznachnosti (anghl. word sense disambiguation ili WSD) — zadacha opriedielieniia smysla (znachieniia) slova, kotoroie prinimaietsia v opriedieliennom kontiekstie [2].

V dannom issliedovanii rassmatrivaietsia mietod postroieniia lieksichieskikh tsiepochiek. Dannyi mietod podrazumievaiet, chto u otryvkov iz razghovornogho ili pis'miennogho tieksta iest' svoistvo iedinstva. Sintaksichieskiie i lieksichieskiie sriedstva moghut ispol'zovat'sia, chtoby sozdat' oshchushchieniie sviaznosti miezhdu priedlozhieniiami, iavlieniie, izviestnoie kak tiekstovaia sviaznost'[4]. Iz vsiekh sriedstv sviazi lieksichieskaia sviaznost', vieroiatno, naibolieie poddaiushchaiasia avtomatichieskoi idientifikatsii. Lieksichieskaia sviaznost' voznikaiet, koghda slova sviazany siemantichieski, naprimier v otnoshieniiakh povtorieniia miezhdu

tierminom i sinonimom. Formirovaniie lieksichieskoi tsiepochki – eto protsiess soiedinieniia siemantichieski sviazannykh slov [3].

V stat'ie [1] s tsiel'iu riefierirovaniia tieksta stroitsia modiel' v vidie lieksichieskikh tsiepochiek. Riefierirovaniie vkliuchaiet chietyrie etapa: orighinal'nyi tiekst dielitsia na bloki (sieghmienty), stroiatsia lieksichieskiie tsiepochki, opriedieliaiutsia sil'nyie tsiepochki, izvliekaiutsia vazhnyie priedlozhieniia.

Sut' mietoda zakliuchaietsia v obiedinienii raznykh chastiei tieksta v odno tsieloie, v to, chto imieiet obshchieie znachieniie (smysl).

V dannoi stat'ie obiediniaiutsia razlichnyie slova v tiekstie s tsiel'iu nakhozhdieniia obshchiegho znachieniia miezhdu nimi. Takim obrazom, proiskhodit izbavlieniie ot lieksichieskoi mnoghoznachnosti.

V stat'ie [4] opisyvaietsia dva sposoba formirovaniia lieksichieskoi sviaznosti:

- Lieksichieskaia sviaznost' povtorienii (reiteration category) dostighaietsia povtorom slov, ispol'zovaniiem sinonimov i ghiponimov;
- Lieksichieskaia sviaznost' slovosochietanii (collocation category) opriedieliena dlia slov, kotoryie chasto upotriebliaiutsia vmiestie, to iest' vstriechaiutsia v odnikh i tiekh zhie kontiekstakh;

Slova i frazy, miezhdu kotorymi sushchiestvuiet lieksichieskaia sviaznost', priedstavliaiut soboi lieksichieskuiu tsiepochku (lexical chains). Mietod lieksichieskikh tsiepochiek osnovan na analizie sovmiestnoi vstriechaiemosti slov i lieksichieskikh sviaziei miezhdu slovami.

Dostoinstvo lieksichieskikh tsiepochiek sostoit v tom, chto ikh nie slozhno raspoznat' i postroit'.

Mietod postroieniia lieksichieskikh tsiepochiek vkliuchaiet shaghi:

1. Vybiraietsia nabor slov-kandidatov (sushchiestvitiel'nyie i sostavnyie sushchiestvitiel'nyie). Eto kandidaty na vkliuchieniie v tsiepochki;
2. Stroitsia spisok vsiekh znachienii dlia kazhdogho slova-kandidata (po dannym slovaria);
3. Dlia kazhdogho znachieniia kazhdogho slova-kandidata nakhoditsia sviaz' dlia kazhdogho slova vo vsiekh uzhie postroiennykh tsiepochkakh (slovo v tsiepochkie imieiet strogho opriedielionnoie znachieniie, zadavaiemyie drughimi slovami v toi zhie tsiepochkie);
4. Slovo-kandidat dobavliaietsia v tsiepochki so slovami, v kotorykh naidiena sviaz'. Smyslovaia nieodnoznachnost' ustraniaietsia, to iest' v tsiepochku dobavliaietsia nie prosto slovo, a iegho konkrietnoie znachieniie;

Dlia illiustratsii mietoda priviediem primier na otryvkie tieksta, priedstavliennogho nizhie, i opriedielim, kakiie znachieniia budut vybrany dlia slov «*любовь*» и «*дом*». Первым существительным в тексте является слово «*любовь*», исходя из данных Русского Викисловаря, у него есть семь различных значений [8]:

1. **чувство** глубокой привязанности к кому-либо, чему-либо; Материнская любовь; Любовь к другу; цит. Люблю **отчизну** я, но странною любовью! Не победит её рассудок мой. [М. Ю. Лермонтов, «Родина», 1841 г.];
2. **чувство** расположения, симпатии к кому-либо;
3. **чувство** горячей сердечной склонности, влечение к другому человеку;
4. чья-то о человеке, внушающем **чувство** любви (в предыдущем значении);
5. любовные отношения;
6. внутреннее стремление, влечение, склонность, тяготение к чему-либо;
7. пристрастие к чему-либо, предпочтение чего-либо;

Наличие нескольких значений разбивает пространство цепочек на несколько множеств интерпретаций, в каждой из которых используются разные значения слова «любовь». Четыре первых значения слова «*любовь*» связаны со словом «*чувство*» и только в первом значении «*любовь*» (цит. Люблю отчизну я, но странною любовью! Не победит её рассудок мой. [М. Ю. Лермонтов, «Родина», 1841 г.]) связана со словом «*Родина*», отсюда получаем две интерпретации (Рис.1).

> **Любовь** к **Родине** – odno iz samykh moshchnykh, vozvyshiennykh **chuvstv**. Ona v polnoi **mierie** proiavilas' v bratskoi **poddierzhkie zhitieliei** Kryma i Sievastopolia, koghda oni tvierdo rieshili viernut'sia v svoi rodnoi **dom**. (V. V. Putin)

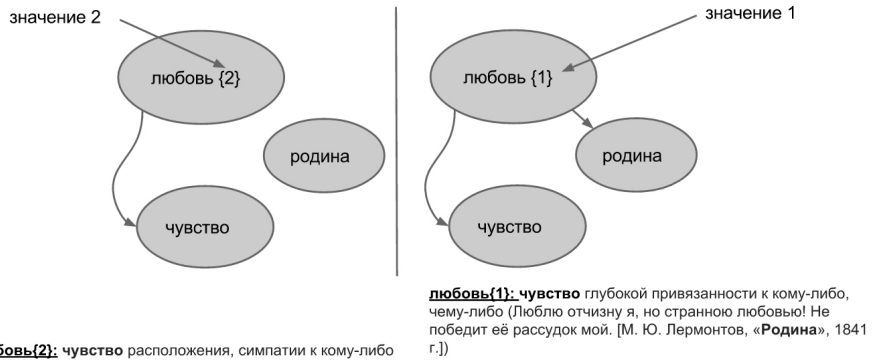

Figure 1: Шаг 1, интерпретация 1 и 2

Компонентой в работе [1] называют список взаимоисключающих интерпретаций. Именно посредством компонент выбор одного из значений слов

ведёт к выбору соответствующей интерпретации, а, следовательно, к невозможности других интерпретаций из этой компоненты. Интерпретации 1 и 2 (Рис. 1) являются компонентой. Следующее слово «*мера*» не связано со словами из первой компоненты [9], поэтому для него создается компонента с одним значением (то есть новая компонента содержит ровно одну интерпретацию). Следующее слово «*поддержка*» также не связано со словами из первой компоненты [11], поэтому для него создается новая компонента с одним значением. Слово «*житель*» имеет единственное значение в Викисловаре [10]:

1. представитель населения; тот, кто живёт где-либо, в чём-либо;

Несмотря на единственное значение у слова «*житель*», есть еще и единственный гипоним «*гражданин*», который имеет значения [6]:

1. лицо мужского рода, принадлежащее к населению какого-либо государства, пользующееся всеми правами и исполняющее все обязанности, установленные законами государства;

2. человек, служащий **родине**, народу, обществу, заботящийся об общественном благе;

3. официальное обращение к мужчине;

Исходя из значений слова «*гражданин*», можно сделать вывод, что слово «*житель*» связано со словом «*родина*» через гипоним «*гражданин*», так как во втором толковании гипонима встречается слово «*родина*».

Таким образом, получается вторая компонента (Рис. 2). Если продолжить этот процесс и добавить слово «*дом*», имеющее семь значений [7], то количество альтернативных вариантов значительно увеличивается. Во втором толковании слова «*дом*» есть слово «*место*», которое можно связать со словом «родина», так как в единственном толковании «*родины*» есть слово «*место*». Также во втором толковании слова «*дом*» есть слово «*проживать*», если мы посмотрим значения этого слова в Викисловаре, то увидим, что первое и второе толкования содержат слово «*жить*», которое можно связать со словом «*житель*». Таким образом получается третья компонента (Рис. 3).

Samyie sil'nyie intierprietatsii priedstavlieny na risunkie. Pri uslovii, chto tiekst sviaznyi, luchshiei intierprietatsiiei schitaietsia ta, kotoraia imieiet bol'shie vsiegho sviaziei (Ris. 4). V dannom sluchaie v kontsie shagha 3 vybrany slieduiushchiie intierprietatsii intieriesuiushchikh nas slov:

- *liubov'* [lieksiema «*любовь*», **значение**: чувство глубокой привязанности к кому-либо, чему-либо];

- *дом* [лексема «*дом*», **значение**: место, где кто-либо постоянно проживает];

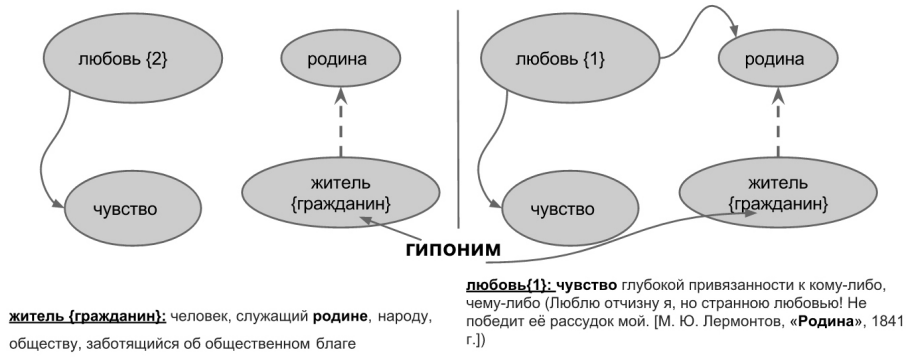

Figure 2: Шаг 2, интерпретации 1 – 2

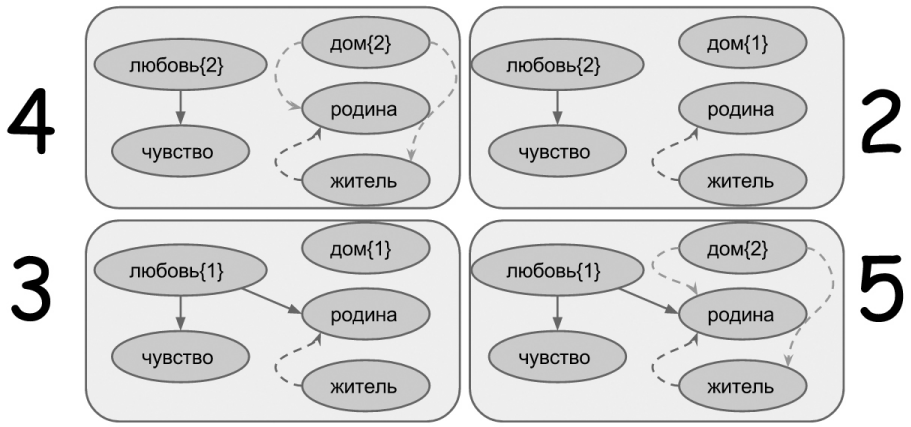

Figure 3: Шаг 3, интерпретации 1 – 4 (Kollichiestvo sviaziei na risunkie uazano bol'shimi tsiframi 4,2,3,5)

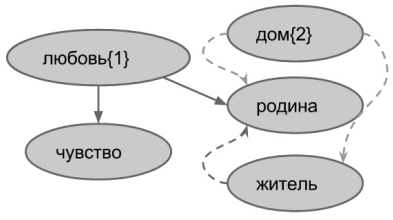

Figure 4: Окончательная лексическая цепочка, полученная путем выбора самой сильной интерпретации

Полученый результат верно отражает значения слов в рассматриваемом контексте.

В данной статье рассмотрен метод построения лексических цепочек для решения задачи WSD. В ходе исследования был разработан алгоритм построения лексической цепочки с помощью Русского Викисловаря. Было показано как словарные статьи Викисловаря могут использоваться в процессе построения лексических цепей.

References

[1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. pages 10–17, 1997.

[2] P. Edmonds and E. Agirre. Word sense disambiguation. 2008.

[3] M. Galley and K. McKeown. Improving word sense disambiguation in lexical chaining. 2003.

[4] M. Halliday and R. Hasan. *Cohesion in English*. Informa UK Limited, 1976.

[5] R. Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. 2007.

[6] Wiktionary. citizen.

[7] Wiktionary. home.

[8] Wiktionary. love.

[9] Wiktionary. measure.

[10] Wiktionary. resident.

[11] Wiktionary. support.