УДК 004.89

Лиман К.С. м.н.с.

Моделі тематик в застосуванні до зняття семантичної неоднозначності

Задача визначення значення слова, з множини можливих значень – онтології, відома як зняття семантичної неоднозначності (WSD, Word Sense Disambiguation). Значення конкретного слова у тексті залежить від його контексту. Значною перевагою використання онтологій, є те, що в них описано відношення між поняттями.

В тематичних моделях, таких як LDA (Latent Dirichlet Allocation), документ представляється як вектор пропорцій тематик, а кожна тематика як розподіл над словами. Але для потреб WSD необхідно мати розподіл над значеннями слів, як наприклад в LDAWN (розширення LDA на WordNet). Але LDAWN не враховує значення семантичної схожості і пов'язаності між концептами. Тому було запропоновано тематичну модель пов'язаних концептів, одним з параметрів якої є взаємне «розташування» концептів задане матрицею пов'язаності.

Ключові слова: семантична неоднозначность, тематичні моделі, семантична схожість.

[1] Міжнародний науково-навчальний центр інформаційних технологій і систем НАН України та МОН України, 03680 м.Київ Глушкова 40, e-mail: lyman.kostiantyn@gmail.com

Lyman K.S. researcher

Topic models application to word sense disambiguty

The problem of choosing a sense for a particular word from fixed set – ontology, is known as Word Sense Disambiguation (WSD). The main advantage of ontologies is that they contain information about semantic relationships among senses, although meaning of a word in a text depends on its context.

Within topic models, such as LDA(Latent Dirichlet Allocation), a document is presented as a vector of topic proportions, and each topic is a distribution over words. But for WSD purposes the distribution over senses is required, as for example in LDAWN (an extension of LDA over WordNet). Unfortunately, LDAWN does not taking into account a sematic similarity and relatedness estimations between concepts. That's why a topic model of related concepts was introduced, which has a matrix of relatedness as one of its parameters.

Key Words: word sense disambiguation, topic models, semantic similarity.

[1] International Research and Training Center for Information Technologies and Systems 03680, Kyiv, Glushkova st. 40, e-mail: lyman.kostiantyn@gmail.com

Статтю представив чл.-кор. НАНУ, д.ф.-м.н., проф. Анісімов А.В.

Зняття семантичної неоднозначності (Word Sense Disambiguation, WSD) є класичною задачею комп'ютерної лінгвістики. Вона полягає в тому, що комп'ютерна система повинна вибрати з запропонованих правильне значення слова в даному контексті. Питання про те якою має бути множина цих значень, наскільки повною та детальною, і те як вона представляється є, напевне, одними з найпроблематичніших у сфері комп'ютерного семантичного аналізу та

інжинірингу знань. Проблема в тому, що люди, хоча і розуміють один одного, не зважаючи на різний мовленнєвий досвід, все таки не завжди відчувають різницю між різними значеннями слова[1]. То чи варто очікувати, що комп'ютер зможе краще? Виділяють два основні підходи до визначення значення слів [1]. Перший, назвемо його *дистрибутивним*, базується на факті, що смисл слова визначається його контекстом. Відомий приклад – вірш Jabberwocky Льюїса Керрола[2].

Інший підхід використовує задані наперед множини можливих значень слів, такі як словники, онтології чи бази знань, створені експертами які і визначають структуру та рівень деталізації семантики слів. Онтології зазвичай, не покривають всі можливі смисли та їх варіації, вони обмежуються або спеціальним доменом знань або деякою частиною загальновживаної лексики (як наприклад WORDNET [8]). Назвемо цей підхід *знання-залежним* (knowledge based). Використання таких джерел структурованої семантики має ту перевагу, що прописані взаємозв'язки між поняттями відображають певні знання експерта про предметну область, але їх створення та підтримання в актуальному стані є трудомісткою задачею.

Задача автоматичного визначення всіх можливих смислів кожного слова відома як Word Sense Induction, WSI [6]. Вона тісно пов'язана з WSD, але відрізняється від неї: в WSD ми знаємо множину понять і нашою задачею є обрати такий елемент цієї множини, що найкраще пасує до контексту. Натомість WSI представляється як задача розділення контексту цільового слова на кластери, кожен з яких відповідає певному смислу цього слова. Приклади таких систем описані в [10] та [6].

Моделі тематик (наприклад, Latent Dirichlet Allocation [4]) є генеративними імовірнісними моделями колекції документів D, тобто вони описують імовірнісний процес генерації спостережуваної колекції документів (задаючи розподіл $p(x|y)$, де x – спостереження, а y – приховані змінні): кожен документ є вектором пропорцій у просторі тематик отриманим з певного тематичного розподілу; тематики при цьому задаються певним розподілом у просторі всіх слів. Інформація закладена у тому яким чином слова сумісно зустрічаються дають змогу знайти такі параметри моделі, що отримані тематичні розподіли можна інтерпретувати саме як тематики. В останні роки з'являються багато робіт покликаних об'єднати WSD та моделі тематик. В цій статті буде розглянуто та проаналізовано одну з таких моделей - LDAWN [5], в якій тематичний розподіл включає розподіл над прихованими змінними, що відповідають синсетам WordNet. В [7], наприклад, розподіл над синсетами не залежить від тематики і слова генеруються або з тематики, або з синсета.

Більш детальний огляд сучасного стану справ в області зняття семантичної неоднозначності, а також огляд систем, підходів та результатів можна знайти в наступних оглядових роботах [12], [13] та збірнику [1].

LDAWN: Тематично-залежне випадкове блукання по онтології

В LDAWN тематика описується не поліноміальним розподілом над елементами словника, як в LDA, а випадковим блуканням по WordNet – WORDNET-WALK . Основною структурною одиницею WordNet є синсет – множина синонімів, об'єднаних певним значенням. Кожен синсет пов'язаний з іншим певними семантичними та лексичними відношеннями.

WORDNET-WALK це імовірнісний процес генерації слів, що використовує для цього гіпонімічні зв'язки[3] між синсетами у WordNet. Згідно [5], уявімо агента, що починає рух з синсету s_0 `entity` – кореня ієрархії іменників (організовану у вигляді таксономії). Агент обирає наступний вузол в своєму випадковому блуканні серед гіпонімів поточного концепту. Дана процедура повторюється поки не буде досягнуто листкового вузла, який відповідає певному слову. Вибір наступного синсету в блуканні визначається відповідним розподілом над множиною синсетів, яка є множиною прихованих випадкових змінних, розподіл яких впливає на

[1] Особливо враховуючи, те що погоджуються люди між собою теж не так часто: в експерименті з двома анотаторами, які виконували розмітку текстів значеннями з WORDNET, згода була досягнута лише в 72.5% випадках [13]

[2] У перекладі Інни Коваль (2004):

Шов печір. Яштпорки слибкі
В трамичці дзигали якраз
Свистали свердлов'ї бідкі
І звинки мрюкали весь час.

[3] Гіпоніми певного поняття є конкретизація ми цього поняття

процес генерації послідовності слів $p\left(\boldsymbol{w}^d \mid \boldsymbol{z}^d, \boldsymbol{\Lambda}^d\right)$.

Для позначення змінних, які приймають значення зі скінченної множини, будемо використовувати нотацію фіктивних змінних:

$$b \in \mathrm{A}: b = \mathrm{A}_a \Leftrightarrow b = a \Leftrightarrow b = \left(0_1, \ldots, 1_a, \ldots, 0_A\right) \quad (1)$$

де $A = \left|\mathrm{A}\right|$. Також замість $b \in \mathrm{A}$ іноді будемо писати $b \in 1:A$, при цьому запис $b = 1:A$ означає, що b послідовно приймає всі значення з A. Введемо наступні позначення:

- $\boldsymbol{w}^d = \left(w_1^d, \ldots, w_{N_d}^d\right)$ – вектор випадкових величин, що приймають значення зі словника V, і представляють документ розміром у N_d слів, $V = \left|\mathrm{V}\right|$;
- $\boldsymbol{z}^d = \left(z_1^d, \ldots, z_{N_d}^d\right)$ – вектор випадкових величин, значення яких вказують на одну з обраних для кожного слова тематик $z_i^d \in \mathrm{K}$, $K = \left|\mathrm{K}\right|$;
- $\boldsymbol{\Lambda}^d = \left(\Lambda_1^d, \ldots, \Lambda_{N_d}^d\right)$ – вектор шляхів від кореня онтології до синсета з якого було згенеровано відповідне слово. Кожен шлях Λ_i^d отримано з відповідного тематичного блукання, обраного згідно значення z_i^d.

Множину всіх синсетів позначимо S, $S = \left|\mathrm{S}\right|$.

LDAWN провадить наступний генеративний процес:

1) Для кожної тематики $t = 1:K$:
   a. Для кожного синсету $s \in \mathrm{S}$, обираємо розподіл переміщень $\boldsymbol{\beta}_s^t \sim \mathrm{Dir}^{c(s)}\left(\boldsymbol{\tau}_s\right)$ – імовірність переходу від s до i-того гіпоніма c_i^s.
2) Для кожного документу $\boldsymbol{w}^d \in \mathrm{D}$, $d = 1:D$:
   a. Обираємо тематичний розподіл $\boldsymbol{\theta}^d \sim \mathrm{Dir}^{\mathrm{K}}\left(\boldsymbol{\alpha}\right)$;
   b. Для кожного слова w_u^d, $u = 1:N_d$:
      i. Обираємо тематику $z_u^d \sim \mathrm{Mult}^K\left(1, \boldsymbol{\theta}^d\right)$;
      ii. Будуємо шлях Λ_n^d який починається в кореневому вузлі λ_0.
      iii. Серед нащадків λ_i:
         1. Обираємо наступний вузол у блуканні $\lambda_{i+1} \sim \mathrm{Mult}^{\mathrm{c}(\lambda_i)}\left(1, \boldsymbol{\beta}_{\lambda_i}^{z_n^d}\right)$;
         2. Якщо λ_{i+1} є листковим вузлом, то генеруємо відповідне слово. Інакше повторюємо.

де $\mathrm{Dir}^X\left(\boldsymbol{\alpha}\right)$ – X-вимірний розподіл Діріхле, $\mathrm{Mult}^{\mathrm{X}}(1, \boldsymbol{\theta})$ – X-значний поліноміальний розподіл з однією спробою, а $c\left(s\right)$. – множина нащадків (гіпонімів) синсету s.

Прихованими змінними цієї моделі є параметри K різних тематичних випадкових блукань WORDNET-WALK $\boldsymbol{\beta}_{1:K}^t$, вектор пропорції тематик у документі $\boldsymbol{\theta}^d$, вектор змінних $\boldsymbol{z}^d$, що визначають тематику для кожного слова, а також шлях у онтології для кожного слова $\boldsymbol{\Lambda}^d$.

Для даної колекції документів $\boldsymbol{w}_{1:D}$ повна апостеріорна імовірність прихованих параметрів:

$$p\left(\boldsymbol{\beta}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\Lambda}_{1:D} \mid \boldsymbol{w}_{1:D}, \boldsymbol{\alpha}, \boldsymbol{\tau}\right) \propto$$
$$\propto \prod_{t=1}^{K} p\left(\boldsymbol{\beta}_t \mid \boldsymbol{\tau}\right) \prod_{d=1}^{D} p\left(\boldsymbol{\theta}_d \mid \alpha\right) \times$$
$$\times \prod_{u=1}^{N_d} p\left(w_u^d \mid \Lambda_u^d\right) p\left(\Lambda_u^d \mid \boldsymbol{\beta}_{z_u^d}\right) p\left(z_u^d \mid \boldsymbol{\theta}_d\right) \quad (2).$$

Тепер задачу визначення значення слова можна поставити наступним чином: яка імовірність, що слово w_u^d було отримано зі шляху, що закінчується синсетом s. Тобто потрібно отримати апостеріорний розподіл прихованих змінних.

Інтуїтивне припущення, на якому базується LDAWN, полягає в тому, що слова однієї тематики семантично близькі і тому будуть поділяти шляхи в WORDNET. При спостереженні певного синсета s (збільшується імовірність спостерігати далі синсети, чиї геренеруючі шляхи мають спільні вузли з аналогічним шляхом для s, тобто мають спільних пращурів.

Нехай апріорне значення $\tau_{s,c}$ імовірності переходу від синсету s до його нащадка c є пропорційним загальній кількості спостережуваних слів, які є нащадками c [5]. В цьому випадку отримаємо імовірнісну інтерпретацію поняття ІНФОРМАЦІЙНОГО КОНТЕНТУ [14].

$$IC(s) = -\ln p(s) \qquad (3)$$

Де $p(s)$ імовірність зустріти s або одного з його нащадків. Цю величину можна оцінити наступним чином

$$\hat{p}(s) = \frac{\left|w \in d : \text{cover}(s, w)\right|}{N_d} \qquad (4)$$

тобто яка частина слів документу D покривається синсетом s (слово w покривається синсетом s, $cover(w,s) = True$, якщо w входить до s або будь-якого його нащадка).

Апостеріорний розподіл для LDAWN

Обчислення апостеріорного розподілу прихованих параметрів пов'язано з інтегруванням по всіх можливих значеннях цих параметрів (для обчислення граничного розподілу спостережуваних даних). Тому на практиці використовуються методи Monte Carlo Markov Chains (MCMC). Одним з них є Гіббсівське семплювання[4]. Для записати умовну імовірність одного елементу вектору випадкових величин, від фіксованих значень решти елементів, таким чином одночасно переміщуючись тільки в одному напрямку.

В LDAWN для кожного слова w_u^d в кожному документі семплюється призначення тематики z_u^d і тематично-залежний шлях через WordNet Λ_u^d. Для обох змінних потрібно обчислити умовний розподіл від спостережуваних даних. Призначення тематики та шляху семплються спільно, при умовній залежності від значень прихованих змінних обраних для всіх інших слів. Запишемо апостеріорний розподіл тематик і шляхів:

$$p\left(z_u^d = t, \Lambda_u^d = \pi \mid \boldsymbol{z}_{-u}^d, \boldsymbol{\Lambda}_{-u}^d, \boldsymbol{\tau}, \boldsymbol{\alpha}\right) = \underbrace{p\left(z_u^d = t \mid \boldsymbol{z}_{-u}^d, \boldsymbol{\alpha}\right)}_{\text{тематика}} \underbrace{p\left(\Lambda_u^d = \pi \mid z_u^d = t, \boldsymbol{\Lambda}_{-u}^{d,t}, \boldsymbol{\tau}\right)}_{\text{шлях}} \qquad (5)$$

Позначимо $n_{-u}^{d,t}$ кількість слів відмінних від u, яким було призначено тематику t в документі d, якому слово w_u належить. Також, нехай $b_{-u}^{t,s,c}$ позначає кількість переходів в випадковому блуканні t-тої тематики від синсету s до його нащадка c, не враховуючи при цьому шлях π пов'язаний зі словом u, тобто кількість переходів від s до c у шляхах $\boldsymbol{\Lambda}_{-u}^{d,t}$. Розглянемо перший множник (імовірність тематики) спільного розподілу:

$$\begin{aligned} &p\left(z_u^d = t \mid \boldsymbol{z}_{-u}^d, \boldsymbol{\alpha}\right) = \\ &= \int_\theta p_{Mult}\left(z_u^d = t \mid \boldsymbol{\theta}\right) p_{Dir}\left(\boldsymbol{\theta} \mid \boldsymbol{z}_{-u}^d, \boldsymbol{\alpha}\right) d\boldsymbol{\theta} = \\ &= \int_\theta \theta_t \frac{\Gamma\left(\sum_{k=1}^K \left[n_{-u}^{d,k} + \alpha_k\right]\right)}{\prod_{k=1}^K \Gamma\left(n_{-u}^{d,k} + \alpha_k\right)} \prod_{k=1}^K \theta_k^{n_{-u}^{d,k} + \alpha_k - 1} d\theta = \\ &= \frac{\Gamma\left(\sum_{k=1}^K \left(n_{-u}^{d,k} + \alpha_k\right)\right)}{\prod_{k=1}^K \Gamma\left(n_{-u}^{d,k} + \alpha_k\right)} \int_\theta \theta_t \prod_{k=1}^K \theta_k^{n_{-u}^{d,k} + \alpha_k - 1} d\theta \end{aligned} \qquad (6)$$

Нехай $\tilde{n}_{d,k} = n_{-u}^{d,k}$ для $k \neq t$ і $\tilde{n}_{d,k} = n_{-u}^{d,k} + 1$ для $k = t$. Тоді, враховуючи рівність $\int_\theta \prod_{k=1}^K \theta_k^{\alpha_k - 1} = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$ отриману з визначення розподілу Діріхле і властивість Гамма-функції $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, отримуємо вираз (7) для імовірності тематики.

$$\begin{aligned} &p\left(z_u^d = t \mid \boldsymbol{z}_{-u}^d, \boldsymbol{\alpha}\right) = \\ &= \frac{\Gamma\left(\sum_{k=1}^K \left[n_{-u}^{d,k} + \alpha_k\right]\right)}{\prod_{k=1}^K \Gamma\left(n_{-u}^{d,k} + \alpha_k\right)} \times \frac{\prod_{k=1}^K \Gamma\left(\tilde{n}_{d,k} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K \left[\tilde{n}_{d,k} + \alpha_k\right]\right)} = \\ &= \frac{n_{-u}^{d,t} + \alpha_t}{\sum_{k=1}^K \left[n_{-u}^{d,k} + \alpha_k\right]} \end{aligned} \qquad (7)$$

Аналогічним чином, розкривається другий множник. Зазначимо, що в цьому випадку доречніше було б писати $\boldsymbol{\Lambda}_{-\pi}^d$ замість $\boldsymbol{\Lambda}_{-u}^d$, щоб

[4] Семплювання – процес отримання зразків деякої випадкової величини.

підкреслити те, що беремо всі шляхи окрім π – шляху, що веде від вершини онтології до слова w_u, але для однорідності нотації продовжимо використовувати Λ_{-u}^d, пам'ятаючи про альтернативне написання. Запишемо (8) – імовірність шляху $\pi = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$, що дорівнює добутку по всім переходам шляху, $|\pi| = L$, $\lambda_i \in \mathrm{S}$:

$$p\left(\Lambda_u^d = \pi \mid z_u^d = t, \Lambda_{-u}^d, \tau\right) = \prod_{l=0}^{L-1} \frac{b_{-u}^{t,\lambda_i,\lambda_{i+1}} + \tau_{\lambda_i,\lambda_{i+1}}}{\sum_{c=1}^{C_{\lambda_i}} \left[b_{-u}^{t,\lambda_i,c} + \tau_{\lambda_i,c} \right]} \quad (8)$$

Отже, отримаємо спільний розподіл призначення тематики і вибору шляху за умови зафіксованих тематик і шляхів для решти слів документу:

$$p\left(z_d^u = t, \Lambda_d^u = \pi | \boldsymbol{z}_d^{-u}, \boldsymbol{\Lambda}_d^{-u}, \tau, \alpha\right) \propto \frac{n_{-u}^{d,t} + \alpha_t}{\sum_{k=1}^{K} \left[n_{-u}^{d,k} + \alpha_k \right]} \left(\prod_{l=0}^{L-1} \frac{b_{-u}^{t,\lambda_i,\lambda_{i+1}} + \tau_{\lambda_i,\lambda_{i+1}}}{\sum_{c=1}^{C_{\lambda_i}} \left[b_{-u}^{t,\lambda_i,\lambda_{i+1}} + \tau_{\lambda_i,c} \right]} \right) \quad (9)$$

Схожість та пов'язаність

Як було сказано вище, для гіперпараметрів τ апріорними значеннями обираються значення їх інформаційного контенту, що у [14] було використано для визначення семантичної схожості синсетів наступним чином:

$$sim_{RES}(s,c) = IC\left(LCS(s,c)\right)$$

де $LCS(s,c)$ – найближчий спільній родовий вузол синсетів s та c, тобто такий предок s та c, що має мінімальну кількість вузлів у шляхах від нього до s та c, відповідно. Інший приклад [10]:

$$sim_{LIN}(s,c) = \frac{2 \times IC\left(LCS(s,c)\right)}{IC(s) + IC(c)}$$

Leacock Claudia та Chodorow Martin в [8] застосували ідею, що чим більша відстань між двома поняттями на графі онтології, тим менш вони семантично схожі:

$$sim_{LCH}(s,c) = -\log \frac{minPath(s,c)}{2D}$$

де D – діаметр таксономії, а $minPath(s,c)$ – довжина найкоротшого шляху між s та c.

Третій підхід до визначення семантичної схожості було запропоновано у [9], який базується на простій ідеї: чим більше спільних слів у словниковому описі двох понять, тим більше вони семантично схожі. Деякі результати порівняння цих методів і суджень людей про схожість слів можна знайти в [2].

Центральною проблемою з оцінкою семантичної схожості двох понять чи слів є визначення того, що ми розуміємо під схожістю. В літературі часто виділять два поняття: схожості і пов'язаності. Два концепти можуть бути просто семантично пов'язаними, наприклад «двигун» і «коробка передач», тобто зустрівши одне з цих понять, зустріч іншого не буде несподіванкою. Але при цьому не можна сказати, що «двигун» і «коробка передач» у чомусь схожі, принаймні не більше ніж «двигун» і «мотор». Остання пара слів, взагалі кажучи є синонімами і це нам вказує на основну властивість схожості: схожі сутності в деякому контексті взаємозамінні . Зауважимо, для наведених мір «схожості» важко визначити чим вони є насправді мірами схожості чи пов'язаності, тому будемо розглядати вихід цих методів як оцінку семантичної пов'язаності.

Уявімо документ не як набір слів $\boldsymbol{w}^d = \left(w_1^d, \ldots, w_N^d\right)$, $w_i^d \in \mathrm{V}$, а як набір синсетів $\boldsymbol{c}^d = (c_1^d, \ldots, c_N^d)$, $c_i^d \in \mathrm{S}$. Тобто, у випадку LDAWN генеративний процес зупиняється, не генеруючи слів.

Тоді, виходячи з попередніх аргументів, семантична пов'язаність двох синсетів s_m та s_h пропорційна імовірності їх одночасної появи у певному контексті. Тобто, імовірності того, що випадкові величини c_u^d та c_v^d одночасно приймуть значення s_m та s_h в документі d. При цьому розглядаємо лише одну певну тематику t :

$$rel_{d,t}(s_m, s_h) = p\left(c_u^d = s_m, c_v^d = s_h \mid \boldsymbol{c}_{-u,v}^d\right) = \quad (10)$$
$$= p_{LDAWN}\left(\Lambda_u^d = \pi_m \mid \Lambda_{-u}^d, \boldsymbol{z}_u = t\right) \times$$
$$\times p_{LDAWN}\left(\Lambda_v^d = \pi_h \mid \boldsymbol{\Lambda}_{-u,v}^d, \boldsymbol{z}_v = t\right) =$$
$$= \prod_{j=0}^{L_m - 1} p_t\left(\lambda_{j+1}^m \mid \lambda_j^m, \boldsymbol{\Lambda}_{-u}^d\right) \prod_{j=0}^{L_h - 1} p_t\left(\lambda_{j+1}^h \mid \lambda_j^h, \boldsymbol{\Lambda}_{-u,v}^d\right)$$

де, згідно з результатом (8) для апостеріорного розподілу шляхів Λ_u^d :

$$\prod_{j=0}^{L_m-1} p_t\left(\lambda_{j+1}^m \mid \lambda_j^m, \boldsymbol{\Lambda}_{-u}^d\right) = \prod_{j=0}^{L_m-1} \frac{b_{-u,v}^{t,\lambda_j^m,\lambda_{j+1}^m} + \tau_{\lambda_j^m,\lambda_{j+1}^m}}{\sum_{c=1}^{C_{\lambda_j^m}} \left[b_{-u,v}^{t,\lambda_j^m,c} + \tau_{\lambda_j^m,c} \right]} \quad (11)$$

Другий множник (10) розкладається аналогічно. Введемо наступні позначення: $\pi_{m,h} = \pi_m \cap \pi_h = \left\{\lambda_0 = s_0, ..., \lambda_{m,h} = LCS\left(s_m, s_h\right)\right\}$ та $\pi_{m\setminus h} = \pi_m \setminus \pi_h = \left\{\lambda_{m\setminus h} = LCS\left(s_m, s_h\right), ..., \lambda_{L_m} = s_m\right\}$, $\pi_{h\setminus m}$ аналогічно. Тоді:

$$rel_{d,t}\left(s_m, s_h\right) = p\left(\pi_{m,h} \mid \boldsymbol{\Lambda}_{-u,v}^d\right)^2 p\left(\pi_{m\setminus h} \mid \boldsymbol{\Lambda}_{-u}^d\right) p\left(\pi_{h\setminus m} \mid \boldsymbol{\Lambda}_{-u,v}^d\right) \quad (12)$$

Відмітимо, що для отримання граничної пов'язаності s_m та s_h потрібно просумувати по всіх можливих шляхах до цих синсетів та всіх контекстах $d = 1:D$.

Тематична модель пов'язаних концептів

Як було показано вище, апостеріорна імовірність вибору певного синсету $p_t\left(c_u^d = s_m \Leftrightarrow \Lambda_u^d = \pi_m \mid \boldsymbol{\Lambda}_{-u}^d\right)$ в LDAWN, а також сумісний розподіл двох синсетів $p_t\left(c_u^d = s_m, c_v^d = s_h \mid \boldsymbol{\Lambda}_{-u,v}^d\right) = rel_t\left(s_m, s_h\right)$ залежить лише від кількості спостережуваних переходів у тематичному випадковому блуканні між синсетами, що входять до шляхів π_m та π_h.

Наявність чи відсутність сиблінгів синсету s (вузлів таксономії того ж рівня, що і s) не впливає на його розподіл. Тому розглянемо наступну генеративну тематичну модель, в якій буде можливо врахувати інформацію про «відстані» між синсетами, що обчислюються за допомогою мір схожості таких як LIN чи LESK.

Розіб'ємо множину синсетів S на рівні: $\mathrm{S}_\delta = \left\{s \in \mathrm{S} : depth\left(s\right) = \delta, \delta = 0:\Delta\right\}$

Генеративний процес:

1) Для кожної тематики $t \in 1:K$
   a. Для кожного рівня $\delta = 0:\Delta$ обираємо розподіл для множини випадкових величин, що відповідають синсетам, $\boldsymbol{s}_\delta^t \sim \mathrm{N}\left(\boldsymbol{s}_\delta \mid \boldsymbol{\mu}_\delta^t\left(\boldsymbol{\mu}_{\delta-1}^t\right), \boldsymbol{\Sigma}_\delta\right)$, де $\boldsymbol{\Sigma}_\delta$ – матриця коваріацій, чиє апріорне значення відповідає rel_δ – матриці попарних схожостей синсетів рівня δ, а $\boldsymbol{\mu}_\delta$ – вектор середніх значень синсетів рівня δ, що в загальному випадку залежить від $\boldsymbol{\mu}_{\delta-1}$: чим більше $\mu_{\delta-1,s}$ тим більші $\mu_{\delta,c}$, де c є нащадком s : $c \in c\left(s\right)$.
2) Для кожного документу $\boldsymbol{w}^d \in \mathrm{D}$, $d = 1:D$:
   a. Обираємо тематичний розподіл $\boldsymbol{\theta}^d \sim \mathrm{Dir}^K\left(\boldsymbol{\alpha}\right)$.
   b. Обираємо розподіл рівнів $\boldsymbol{G}^d \sim \mathrm{Dir}^\Delta\left(\boldsymbol{\gamma}\right)$.
   c. Для кожного слова w_u^d , $u = 1:N_d$:
      i. Обираємо тематику $z_u^d \sim \mathrm{Mult}^{\mathrm{K}}\left(1, \boldsymbol{\theta}^d\right)$
      ii. Обираємо рівень $g_u^d \sim \mathrm{Mult}^{\Delta}\left(1, \boldsymbol{G}^d\right)$
      iii. Обираємо синсет $c_u^d \sim Mult^S\left(1, \sigma\left(\boldsymbol{s}_{g_u}^{z_u} \mid \boldsymbol{\mu}_{g_u}^{z_u}, \boldsymbol{\Sigma}_{g_u}^{z_u}\right)\right)$
      iv. Генеруємо слово w_u^d з обраного синсету c_u^d .

де $\sigma_i\left(\boldsymbol{\alpha}\right) = \frac{\exp\left(\alpha_i\right)}{\sum_j \exp\left(\alpha_j\right)}$ – softmax функція, яка приймає значення в інтервалі $\left[0;1\right]$. Тобто імовірність, що u-те слово буде згенеровано з m-того синсету s_m:.

$$p\left(c_u^d = s_m \mid \mu_\delta^t, \Sigma_\delta^t\right) = \sigma_m\left(\boldsymbol{s}_\delta^t \mid \boldsymbol{\mu}_\delta^t, \boldsymbol{\Sigma}_\delta^t\right) \quad (13)$$

Повна апостеріорна імовірність прихованих параметрів:

$$p\left(\boldsymbol{\theta}^d, \boldsymbol{G}^d, \boldsymbol{g}^d, \boldsymbol{z}^d, \boldsymbol{c}^d, \boldsymbol{s}^d, \boldsymbol{\mu}^d \mid \boldsymbol{w}^d, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) = \frac{p\left(\boldsymbol{w}^d \mid \boldsymbol{\theta}^d, \boldsymbol{G}^d, \boldsymbol{g}^d, \boldsymbol{z}^d, \boldsymbol{c}^d, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right)}{\int_{\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{g},\boldsymbol{s},\boldsymbol{c},\boldsymbol{z}} p\left(\boldsymbol{w}^d \mid \boldsymbol{\theta}^d, \boldsymbol{G}^d, \boldsymbol{g}^d, \boldsymbol{z}^d, \boldsymbol{c}^d, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right)} \quad (14)$$

Запишемо функцію правдоподібності для документу $\boldsymbol{w}^d$:

$$\begin{aligned} &p\left(\boldsymbol{w}^d \mid \boldsymbol{\theta}^d, \boldsymbol{G}^d, \boldsymbol{g}^d, \boldsymbol{z}^d, \boldsymbol{c}^d, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) = \\ &= p\left(\boldsymbol{\theta}^d \mid \boldsymbol{\alpha}\right) p\left(\boldsymbol{G}^d \mid \boldsymbol{\gamma}\right) \times \\ &\times \prod_{t=1}^{K} p\left(\boldsymbol{s}^t \mid \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right) \prod_{u=1}^{N_d} p\left(w_u^d \mid c_u^d\right) \times \\ &\times p\left(c_u^d \mid \sigma\left(\boldsymbol{s}_{g_u}^{z_u}\right)\right) p\left(g_u^d \mid \boldsymbol{G}^d\right) p\left(z_u^d \mid \boldsymbol{\theta}^d\right) \end{aligned} \quad (15)$$

Вивід апостеріорного розподілу для рівнів $\boldsymbol{g}^d$ і пропорції рівнів $\boldsymbol{G}^d$, схожий на такий для тематик $\boldsymbol{z}^d$ та пропорцій тематик у документі $\boldsymbol{\theta}^d$, бо вони мають однакову структуру розподілу: $p_{\text{Mult}}\left(y \mid \boldsymbol{\theta}\right) p_{\text{Dir}}\left(\boldsymbol{\theta} \mid \boldsymbol{\alpha}\right)$. Для пропорції рівнів легко отримати оцінку максимальної правдоподібності $\hat{G}_\delta = \frac{n_\delta}{N_d}$, де n_δ – кількість спостережуваних синсетів рівня δ.

Розглянемо вивід для розподілу синсетів при фіксованій тематиці t. Нехай ми спостерігаємо документ у вигляді синсетів $\boldsymbol{c}^d = \left(c_1^d, \ldots, c_{N_d^t}^d\right)$, $c_i \in \mathrm{S} : c_i = \mathrm{S}_m \Leftrightarrow c_i = \left(0_1, \ldots, 1_m, \ldots, 0_S\right)$; верхній індекс t в N_d^t використано, щоб підкреслити, що розглядаються лише ті синсети для яких було обрано тематику t, але далі опустимо цей індекс, щоб спростити запис. Згідно опису моделі

$$\begin{aligned} &\boldsymbol{c}^d \sim \text{Mult}\left(N_d, \sigma\left(\boldsymbol{s}\right)\right) = \\ &= \frac{N_d!}{c^{(1)}! c^{(2)}! \ldots c^{(S)}!} p_1^{c^{(1)}} \cdot p_2^{c^{(2)}} \cdot \ldots \cdot p_S^{c^{(S)}} \end{aligned} \quad (16)$$

де $p_m = \sigma\left(s_m\right)$, а $c^{(m)}$ – кількість таких c_u^d, що $c_u^d = s_m$. Пригадаємо, що $\boldsymbol{s} \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ і тоді апостеріорний розподіл синсетів:

$$p\left(\boldsymbol{s} \mid \boldsymbol{c}\right) = \frac{p\left(\boldsymbol{c} \mid \boldsymbol{s}\right) p\left(\boldsymbol{s} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{\int_{\boldsymbol{s}} p\left(\boldsymbol{c} \mid \boldsymbol{s}\right) p\left(\boldsymbol{s} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)} \quad (17)$$

Цей інтеграл не обчислюється аналітично, із-за входження softmax-фнукції, тому апроксимуємо його методом Лапласа нормальним розподілом $q\left(\boldsymbol{s}\right)$. Для цього запишемо логарифмічну функцію правдоподібності:

$$\begin{aligned} &L = -\ln p\left(\boldsymbol{s} \mid \boldsymbol{c}\right) = \\ &= \frac{1}{2}\left(\boldsymbol{s} - \boldsymbol{\mu}\right)^{\mathrm{T}} \Sigma^{-1} \left(\boldsymbol{s} - \boldsymbol{\mu}\right) - \\ &\quad - \sum_{u=1}^{N_d} \sum_{m=1}^{S} c_u^m \ln \sigma\left(s_m\right) + const \end{aligned} \quad (18)$$

Тепер розглянемо першу і другу похідну другого доданку:

$$\begin{aligned} &\nabla_{s_j} \left(\sum_{m=1}^{S} c^{(m)} \ln \sigma\left(s_m\right) \right) = \\ &= -\sum_{m=1}^{S} c^{(m)} \left(I_{m,j} - \sigma\left(s_j\right) \right) \\ &= \begin{cases} N_d \sigma(s_j) & , m \neq j \\ N_d \sigma(s_j) - c^{(m)} & , m = j \end{cases} \end{aligned} \quad (19)$$

$$\begin{aligned} &\nabla_{s_k} \nabla_{s_j} L = \nabla_{s_k} \left(-\sum_{m=1}^{S} c^{(m)} \left(I_{m,j} - \sigma\left(s_j\right) \right) \right) = \\ &= N_d \nabla_{s_k} \sigma\left(s_j\right) = N_d \sigma\left(s_j\right) \left(I_{j,k} - \sigma\left(s_k\right) \right) \end{aligned} \quad (20)$$

Тоді:

$$\nabla\nabla L = \boldsymbol{\Sigma}^{-1} + N_d \left(\sigma\left(\boldsymbol{s}\right) I - \sigma\left(\boldsymbol{s}\right) \sigma^{\mathrm{T}}\left(\boldsymbol{s}\right) \right) \quad (21)$$

Тепер можемо записати апроксимуючий розподіл $q\left(\boldsymbol{s}\right)$ [3]:

$$q\left(\boldsymbol{s}\right) = \mathrm{N}\left(\boldsymbol{s} \mid \boldsymbol{\mu}_{MAP}, \boldsymbol{\Sigma}_N\right) \quad (22)$$

де $\boldsymbol{\Sigma}_N^{-1} = \nabla\nabla L$, а $\boldsymbol{\mu}_{MAP}$ максимальна апостеріорна оцінка, яку можна отримати з ∇L.

Розглянемо тепер дві випадкові величини c_u^d та c_v^d, що відповідають u-тому та v-тому слову і мають значення s_m та s_h відповідно. Нехай для них було обрано однакову тематику $z_u^d = t$, $z_v^d = t$, і обрано рівні $g_u^d = \delta_u$ та $g_v^d = \delta_v$ відповідно. Запишемо вираз для пов'язаності синсетів s_m та s_h, тобто сумісний розподіл

$p_t\left(c_u^d = s_m, c_v^d = s_h \mid \boldsymbol{\mu}^d, \boldsymbol{\Sigma}\right)$, де $\boldsymbol{\mu}^d$ – апостеріорне значення середніх для розподілу синсетів, $\boldsymbol{s}^d \sim \mathrm{N}\left(\boldsymbol{\mu}^d, \boldsymbol{\Sigma}\right)$ після спостереження частини документу $\boldsymbol{w}_{-u,v}^d$, тобто контексту для c_u^d я u та v:

$$rel_{d,t}\left(s_m, s_h\right) = p_t\left(c_u^d = s_m, c_v^d = s_h \mid \boldsymbol{\mu}^d, \boldsymbol{\Sigma}^d\right) \approx \approx \sigma\left(s_m\right)\sigma\left(s_h\right) q\left(\boldsymbol{s}^d \mid \boldsymbol{\mu}^d, \boldsymbol{\Sigma}^d\right) \quad (23)$$

де $\left(\boldsymbol{\Sigma}^d\right)^{-1} = \boldsymbol{\Sigma}^{-1} + \left(N_d - 2\right)\left(\sigma\left(\mathbf{s}\right) I - \sigma\left(\mathbf{s}\right)\sigma^T\left(\mathbf{s}\right)\right)$

Висновок

Таким чином, в запропонованій моделі дистрибутивна пов'язаність прямо залежить від значення структурної пов'язаної – матриці Σ.

Список використаних джерел

[1] *E.Agirre, P.Edmonds* Word Sense Disambiguation: Algorithms and Applications. - Dordrecht, Netherlands : Springer, 2006. - p. 388.

[2] *Anisimov A.V., Lyman K.S., Marchenko O.O.* The computational methods for the semantic proximity measures of natural language words // Artificial Intelligence (Штучний інтелект). - 2010. - 3. - pp. 170-176.

[3] *Christopher M. Bishop* Pattern recognition and Machine learning. - : Springer, 2006.

[4] *D.Blei, A.Ng, M. Jordan* Latent dirichlet allocation // Journal of Machine Learning Research. - : JMLR.org, 2003. - cc. 993-1022.

[5] *J. Boyd-Graber, D.Blei, X. Zhu* A topic model for word sense disambiguation // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). - 2007. - pp. 1024-1033.

[6] *S. Brody, M.Lapata* Bayesian word sense induction // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. - : Association for Computational Linguistics, 2009. - стр. 103-111.

[7] *C.Chemudugunta et al.* Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning // The Semantic Web - ISWC 2008 - Berlin : Springer, 2008. - Vol. 5318.

[8] *C.Fellbaum ed.* WordNet: an electronic lexical database.. - Cambridge, MA : The MIT Press, 1998.

[9] *M.Lesk* Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. // Proceedings of SIGDOC '86.. - 1986. - стр. 24-26.

[10] *D.Lin* Automatic retrieval and clustering of similar words // Proceedings of COLING-ACL 98. - Montreal, Canada :, 1998.

[11] *D.McCarthy* Word Sense Disambiguation: An Overview // Language and Linguistics compass. - : Wiley Online Library, 2009 - 2 : Vol. 3. - стр. 537-558.

[12] *R.Navigli* Word Sense Disambiguation: A Survey // ACM Computing Surveys (CSUR). - : ACM, 2009. - 2 : Vol. 41. - pp. Article 10.

[13] *B.Snyder, M.Palmer* The English all-words task // Proceedings of the ACL senseval-3 workshop. - Barcelona, Spain : ACL, 2004. - стр. 41-43.

[14] *P.Resnik* Using Information Content to Evaluate Semantic Similarity in a Taxonomy // Proceedings of the 14th International Joint Conference on Artificial Intelligence. - 1995. - pp. 448-453.

Надійшла до редколегії 12.02.2013