Simítás hasonlósági információ felhasználásával

Bíró István, Szamonek Zoltán, Szepesvári Csaba

MTA SZTAKI, Gépi Tanulás Kutatócsoport
1111, Budapest, Kende utca 13-17,
e-mail: szcsaba@sztaki.hu, zszami@elte.hu, ibiro@sztaki.hu

1. Bevezetés

Ebben a dolgozatban azt vizsgáljuk, hogy hogyan lehet a szavak egymáshoz való viszonyára vonatkozó információt kihasználva javítani a nyelvmodellek minőségén. Elviekben világos, hogy a szavak disztribúciós hasonlóságát kihasználva ugyannyi adat esetén jobb modelleket lehet építeni. Mivel azonban a disztribúciós hasonlóságra vonatkozó információ nem tökéletes, kérdéses, hogy az ebből adódó hiba ellenére is működhet-e egy a szóhasonlóságokra építő módszer.

A dolgozat fő eredménye az „SBS" (similarity based smoothing) algoritmus, amelyik képes kihasználni a szavakra vonatkozó hasonlósági információt, amennyiben ez az információ kellően pontos, míg ha az információ nem pontos, akkor az algoritmus addícionális vesztesége elhanyagolható.

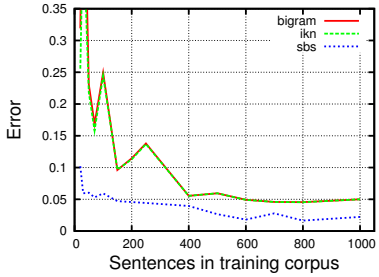
2. A javasolt módszer

A továbbiakban a módszert abban az esetben illusztráljuk, amikor bigram valószínűségeket ($\mathbb{P}(w_k|w_{k-1})$) akarunk modellezni.[1] A kiindulási pont az az észrevétel, hogy az egyes szókettesek valószínűségeit megfelelő ϕ_i bázisfüggvények választásával, a $\log \mathbb{P}(y|x) \propto \sum_{i\in I} \theta_i \phi_i(x,y)$ alakba írhatjuk és a tanulási feladatot így mint egy logisztikus regresszió feladatot is felírhatjuk. Egyéb információ hiányában minden információ-veszteséget elkerülő bázisfüggvényrendszernek „tartalmaznia kell" a $\phi_i(x,y) = \mathbb{I}_{\{i_1=x\}}\mathbb{I}_{\{i_2=y\}}$ függvényeket.[2]

A javasolt módszer lényege, hogy a szavak hasonlóságára vonatkozó információt a bázisfüggvények választásával visszük be az algoritmusba. Illusztrációként tekintsük azt az esetet, amikor a szavak csoportokba sorolhatóak és a $p(y|x)$ valószínűségek csak x csoportjától függ. Nyilvánvaló, hogy ebben az esetben a megfelelő bázisfüggvény választás $\psi_i(x,y) = \mathbb{I}_{\{i_1=y\}}\mathbb{I}_{\{c=C(x)\}}$, ahol $C(x) \in \mathcal{C}$ az x szó csoportja és $i = (i_1, c) \in \mathcal{W} \times \mathcal{C}$. Látható, hogy a bázisfüggvények száma nagyban csökkenthető, ha $\mathcal{C}$ számossága sokkal kisebb, mint $\mathcal{W}$ számossága.

[1] A bonyolultabb modellezési problémákra a bemutatandó elvek alapján triviális a módszert kiterjeszteni.

[2] Itt $i = (i_1, i_2) \in I = \mathcal{W} \times \mathcal{W}$.



1. ábra. A hiba a tanuló adat méretének függvényében.

Amennyiben bizonytalan, hogy a szócsoportok kellő információt tartalmaznake, akkor a fenti $\phi_\cdot$ és az itteni $\psi_\cdot$ bázisfüggvények együttesére kell építeni.Mivel az így kapott rendszer számossága már nagyobb lesz $|\mathcal{W}|^2$-nél, ahhoz, hogy ne veszítsünk a teljesítményből se akkor, ha relevánsak az osztályok, se akkor, ha nem azok, a regressziós irodalomban jól ismert regularizációt javasoljuk használni.Megmutatható, hogy az így kapott módszer valóban akkor is hatékony, amikor a szóosztályok relevánsak és akkor is, amikor nem.

Ha csak szavak közötti hasonlóságok állnak rendelkezésre, akkor a természetes megoldás a bázisfüggvények definiálására az ún. spektrális klaszterezés [1]. Az SBS algoritmus tehát egy a szavak felett definiált hasonlósági mátrixból kiindulva spektrális klaszterezés segítségével meghatározza bázisfüggvények egy rendszerét és erre építve a fent definiált regularizált logisztikus regressziós feladatot megoldva épít sztochasztikus nyelvmodelleket.

Az előzetes kísérletekben a módszer érzékenységét kontrollált környezetben, több szempontból is vizsgáltuk: a hasonlóságra vonatkozó információ minőségére, a módszer paramétereinek beállítására, illetve a rendelkezésre álló adatok mennyiségére nézve. Az 1. ábrán a tanult modell minősége látható a rendelkezésre álló minták számának függvényében.[3] A maximum norma hiba nyilván igen pesszimistán értékeli a modelleket (ezért nem látszik jobbnak az IKN mint a bigram az ábran), így igen biztató, hogy az SBS ebben a normában (is) jelentősen javít a korábbi módszerek eredményein. A valódi adatokkal végzett előzetes kísérletek eredmények szerint a módszer versenyképes a jelenlegi legjobb módszerekkel is.

Hivatkozások

1. Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, pages 269–274, 2001.
2. Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[3] Összeasonlítási alapként a maximum-likelihood becslést ("bigram") és az IKN módszert használtuk. Az IKN [2] az átlagos egy state-of-the-art módszer, amelyiknek szintén az a célja, hogy a kis minták okozta minőségromlási problémákat megoldja.