SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Metoda ispitivanja statističke značajnosti u obradi prirodnog jezika

Filip Boltužić
Voditelj: *Prof. dr. sc. Bojana Dalbelo-Bašić*

Zagreb, rujan 2013.

SADRŽAJ

1. Uvod

U obradi prirodnog jezika (engl. *natural language processing - NLP*) verifikacija izgrađenog sustava (za prevođenje, klasifikaciju govornih činova, izgradnje stabla međuovisnosti ...) nužan je korak. Uz argumentirano obrazloženje autora sustava o dobivenim performansama sustava, potrebno je priložiti statistički dokaz kojime se nepobitno pokazuje kako je uistinu ostvaren doprinos znanstvenoj zajednici. Ispitivanjem statističke značajnosti pokazuje se koliko su rezultati eksperimenata vrijedni, jesu li dobiveni slučajno i u kolikoj mjeri su pouzdani. Ovaj korak je sastavni dio velikog broja NLP članaka.

(Chinchor, 1992) se bavi analizom rezultata MUC-4 (engl. *Message Understanding Conference*), (Koehn, 2004) i (Zhang et al., 2004) ispituju značajnost rezultata strojnog prevođenja (engl. *machine translation*), (Bisani i Ney, 2004) analiziraju rezultate automatskog prepoznavanja govora (engl. *automated speech recognition*). (Berg-Kirkpatrick et al., 2012), (Yeh, 2000) , (Thompson, 1993) se nisu usredotočili na specifičnu metriku (kao što je F1-mjera), već općenitijim metodama ispitivanja statističke pouzdanosti.

U nastavku seminarskog rada opisat će se različiti načini ispitivanja statističke značajnosti iz navedenih radova.

2. Ispitivanje statističke značajnosti

(Thompson, 1993) navodi neke kritike konvencionalnih metoda statističkog testiranja:

1. Nulta hipoteza će se **uvijek** odbaciti ako uzme u obzir dovoljno velika populacija. To je potkrepljeno citatom:

> Ispitivanje statističke značajnosti može biti vođeno tautologijom. Umorni istraživači, nakon prikupljanja skupa podataka, rade statističke testove kako bi se uvjerili da je prikupljena dovoljno velika količina podataka, što je podatak koji već znaju. Ovakva tautologija učinila je mnogo štete znanstvenoj zajednici.

2. Korištenje ANOVE može dovesti do pogrešnih usporedbi. U višedimenzionalnim analizama primjenom hijerarhijskog pristupa moguće je povlačiti usporedbe između podataka iz različitih dimenzija, analogno poslovici: usporedba krušaka i jabuka.

3. Oslanjanje na ispitivanje statističke značajnosti stvara neizbježne dvojbe.

ANOVA zahtjeva spajanje varijanci prilikom izračuna srednje devijacije (u nazivniku). Ova operacija je dozvoljena samo u slučaju da su varijable homogene. Slično tome, ANCOVA (analiza kovarijance) pretpostavlja da je zadovoljen uvjet *homogenosti regresije*.

Bruce Thompson smatra kako ne treba odbaciti ispitivanje statističke značajnosti, već se treba koristiti u prave (skromnije no sada) svrhe.

2.1. Testiranje hipoteze

Dobiveni sustav *A* uspoređuje se s osnovnim (engl. *baseline*) sustavom *B*. Usporedba se radi s nekim dostupnim skupom podataka (engl. *dataset*) - populacijom. Uzimanjem uzorka x iz populacije i usporedbom performansi sustava *A* i *B* nad izabranim uzorkom dobiva se mjera razlike u performansama sustava $\delta(x)$. Testiranjem hipoteze ograđuje se da su dobiveni rezultati slučajnost. Cilj je pokazati kako uzimanjem uzorka $x^{'}$ rezultati (razlike u performansama) će i dalje biti *slični*. Na ovaj način oblikuje se

nulta hipoteza. Nultom hipotezom pretpostavlja se upravo suprotno: ne postoji razlika u performansama između sustava *A* i *B*. Nulta hipoteza označava se sa H_0.

Testiranje hipoteze procjenjuje kolika je to vjerojatnost:

$$p(\delta(X) > \delta(x)|H_0) < \alpha \tag{2.1}$$

gdje je *X* slučajna varijabla mogućih dobivenih uzoraka veličine *n*, a $\delta(x)$ promatrana (konstantna) vrijednost. Ako vrijedi 2.1 za $\alpha = 0.05$ onda se odbacuje nulta hipoteza, jer je 2.1 naziva se *p-vrijednost*, odnosno empirijska razina značajnosti.

P-vrijednost se često aproksimira, jer je u nekim slučajevima nije moguće jednostavno izračunati. Jedna od najčešće korištenih metoda za procjenjivanje p-vrijednosti je upareni (engl. *bootstrap*). Upareni bootstrap jedna je od najčešće korištenih metoda (Koehn, 2004) zato što se može primjeniti na sve mjerne metode (uključujući složenije kao što su BLEU (Papineni et al., 2002), F1).

2.1.1. Bootstrap

Bootstrap procjenjuje *p-vrijednost*, vadi testne uzorke x_i iz populacije i broji koliko često sustav *A* funkcionira s performansama $\delta(x)$ (ili većim) od sustava *B*. Na raspolaganju stoji samo populacija x, pa se podaci iz x uzorkuju sa zamjenom (engl. *sampling with replacement*). Tako dobiveni uzorci nazivaju se (engl. *bootstrap*) uzorcima.

Za dobivene uzorke x_i bi, prema nultoj hipotezi, trebalo vrijediti da

$$\frac{1}{k}\sum_{i=0}^{k}\delta(x_i) = \delta(x) \tag{2.2}$$

gdje je *k* broj uzoraka. Ako se želi provjeriti može li se odbaciti nulta hipoteza, potrebno je provjeriti *koliko često sustav A daje rezultate koji su bolji od očekivanih*. Očekivani rezultat je da je sustav *A* bolji od sustava *B* za $\delta(x)$. Prema tome, prebrojava se u koliko slučajeva test skupova podataka x_i je *A* bio bolji od *B* za $2\delta(x)$. Pseudokod bootstrap postupka je prikazan je (Berg-Kirkpatrick et al., 2012) bootstrap postupka prikazan je 1.

Najveća prednost bootstrap metode je mogućnost računanja $\delta(x)$ za bilo koju metriku.

Algoritam 1 Pseudokod bootstrap postupka

Generiranje b bootstrap uzoraka x_i veličine n nasumičnim izborom s ponavljanjem iz populacije x
$s = 0$
repeat
    if $\delta(x_i) > 2\delta(x)$ **then**
        $s = s + 1$
    end if
until $i > n$
$p = \frac{s}{b}$

3. Zaključak

Zaključak.

4. Literatura

Taylor Berg-Kirkpatrick, David Burkett, i Dan Klein. An empirical investigation of statistical significance in nlp. U *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, stranice 995–1005. Association for Computational Linguistics, 2012.

Maximilian Bisani i Hermann Ney. Bootstrap estimates for confidence intervals in asr performance evaluation. U *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, svezak 1, stranice I–409. IEEE, 2004.

Nancy Chinchor. The statistical significance of the muc-4 results. U *Proceedings of the 4th conference on Message understanding*, stranice 30–50. Association for Computational Linguistics, 1992.

Philipp Koehn. Statistical significance tests for machine translation evaluation. U *EMNLP*, stranice 388–395, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, i Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. U *Proceedings of the 40th annual meeting on association for computational linguistics*, stranice 311–318. Association for Computational Linguistics, 2002.

Bruce Thompson. The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Educational*, stranice 361–377, 1993.

Alexander Yeh. More accurate tests for the statistical significance of result differences. U *Proceedings of the 18th conference on Computational linguistics-Volume 2*, stranice 947–953. Association for Computational Linguistics, 2000.

Ying Zhang, Stephan Vogel, i Alex Waibel. Interpreting bleu/nist scores: How much improvement do we need to have a better system? U *LREC*, 2004.