Improving SMT by Using Multiple Translation Hypotheses

Saša Hasan, Evgeny Matusov, Arne Mauser,
David Vilar, Richard Zens, Hermann Ney

TC-STAR OpenLab on Speech Translation
Trento, Italy – March 30 - April 1, 2006

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University, Germany

Outline

1. Introduction
2. Related work
3. Rescoring
   - Reranking N-best lists
   - Different methods for rescoring
   - Experiments
4. System combination
   - Computing alignments
   - “Voting” on confusion networks
   - Experiments
5. Conclusions

Introduction

- SMT systems (e.g. phrase-based decoders)
  - use a combination of various models during generation
  - are capable of producing single-best output
  - generate word graphs / N-best lists with multiple translation hypotheses
- Observation: all MT systems make errors
- Assumption: different MT systems make different errors (due to utilizing different models / generation strategies / tweaks)
- Two possibilities for improvement:
  - rerank multiple translation candidates from a single MT system
    $\longrightarrow$ *Rescoring*
  - generate consensus translations from various MT systems
    $\longrightarrow$ *System Combination*

Related work

- **Rescoring**
  - **discriminative and minimum error rate training [Och & Ney 02, Och 03]**
  - **different discriminative reranking techniques [Shen & Sarkar$^+$ 04]**
  - **syntactical features for rescoring [Och & Gildea$^+$ 04, Hasan & Bender$^+$ 06]**
  - **clustered language models [Hasan & Ney 05]**
- **System combination**
  - **successful approaches to system combination in automatic speech recognition (ASR) like ROVER [Fiscus 97]**
  - **sentence selection algorithms [Nomoto 04, Paul & Doi$^+$ 05]**
    - **selection of hypotheses based on scores of statistical and other models**
    - **approaches require comparable scores**
  - **algorithms computing consensus translations:**
    - **edit distance based alignment, no reordering [Bangalore & Bordel$^+$ 01]**
    - **heuristic alignment with reordering [Jayaraman & Lavie 05]**

Rescoring

Possible SMT system outputs:

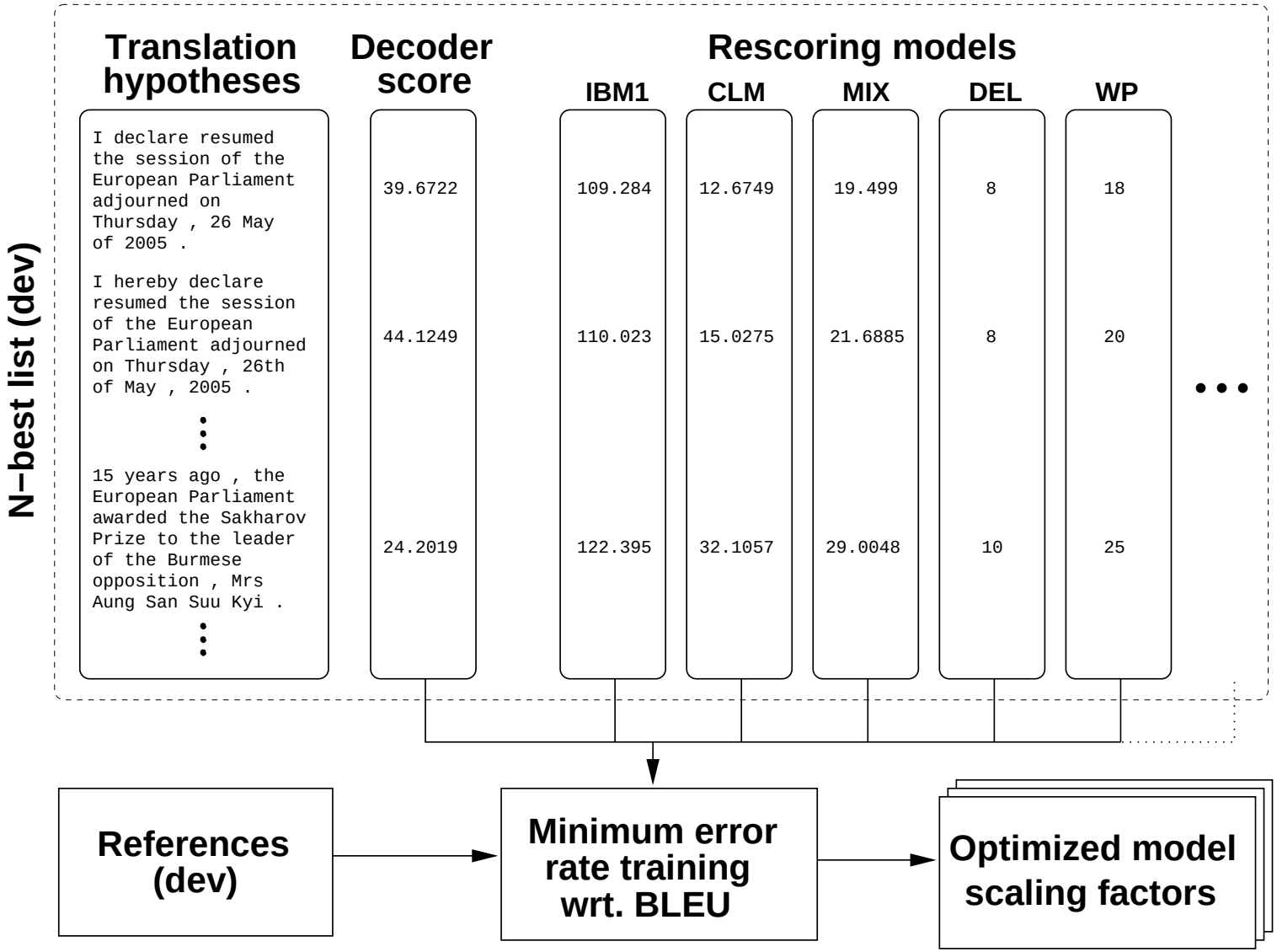
- **single-best (hypothesis with lowest cost / highest probability)**
- **word graph (compact representation of search space): only local rescoring techniques are possible**
- **N-best list (extract of N best hypotheses): rescoring techniques that consider the whole sentence are possible**

Idea of reranking / rescoring:

Reevaluate N-best translation hypotheses of an MT system by adding additional models (features) to the baseline

- **features should be able to distinguish “good” from “bad” translations**
- **discriminatively rerank the translations in a log-linear combination of all models**

Rescoring framework



Rescoring models

- **Syntax-based**
  - **IBM model 1**
  - **using grammars (supertagging, link grammar, parsing)**
  - **ME-based chunking**
- **Language-model based**
  - **high-order n-grams**
  - **sentence-level mixtures**
  - **clustered LMs**
- **Penalties**
  - **IBM1 deletion model**
  - **word / sentence-length penalties**

Applied in a log-linear framework (feature scores denote costs):

$$\hat{e}(f_1^J; \lambda_1^M) = \underset{e_1^I}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

Rescoring models – Details

IBM1:

$$h_{\text{IBM1}}(f_1^J, e_1^I) = -\log\left(\frac{1}{(I+1)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} p(f_j|e_i)\right)$$

Clustered LMs:

$$h_{\text{CLM}}(f_1^J, e_1^I) = -\log\sum_c \left[\, \mathcal{R}_c(f_1^J, e_1^I) \,\right]\left(\alpha_c p_c(e_1^I) + (1-\alpha_c)p_g(e_1^I)\right)$$

Sentence-level mixtures:

$$h_{\text{SLM}}(e_1^I) = -\log\sum_c \mu_c p_c(e_1^I)$$

IBM1 deletion model:

$$h_{\text{Del}}(f_1^J, e_1^I) = \sum_{j=1}^{J}\prod_{i=0}^{I}\left[\, p(f_j|e_i) < \tau \,\right]$$

Model scaling factors

Training criteria for the model scaling factors λ_m, $m = \{1, \ldots, M\}$:

- **Maximum class posterior probability using the GIS algorithm**

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(e_s, f_s) \right\}$$

- **Minimum error rate training using the Downhill Simplex algorithm**

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} E(r_s, \hat{e}(f_s; \lambda_1^M)) \right\}$$

Rescoring experiments

Spanish-English FTE, $N = 10\,000$, optimized wrt. BLEU:

Dev'06	mWER[%]	mPER[%]	BLEU[%]	NIST
Baseline	38.7	27.2	52.0	10.56
+LM	38.6	27.1	52.4	10.59
+IBM	38.5	26.9	52.4	10.62
+IBM+Del	38.5	26.9	52.5	10.62
+IBM+LM	38.3	26.7	52.7	10.67
+IBM+LM+Del	38.2	26.8	52.8	10.67
+IBM+LM+Del+Length	38.2	26.8	52.9	10.66
Oracle (WER, $N = 10k$)	27.3	20.1	64.2	11.91

Eval'06 (official results)	mWER[%]	mPER[%]	BLEU[%]	NIST
Baseline	42.7	31.0	46.6	10.29
+IBM+LM+Del+Length	42.3	30.5	47.7	10.44

Rescoring experiments (contd)

Spanish-English Verbatim, $N = 10\,000$, optimized wrt. BLEU:

Dev'06	mWER[%]	mPER[%]	BLEU[%]	NIST
Baseline	40.4	28.3	51.0	10.43
+LM	40.3	28.3	51.1	10.43
+IBM	39.9	27.8	51.6	10.52
+IBM+Del	39.9	27.9	51.7	10.54
+IBM+LM	39.7	27.7	51.9	10.58
+IBM+LM+Del	39.8	27.8	51.9	10.56
+IBM+LM+Del+Length	39.7	27.7	52.0	10.57
Oracle (WER, $N = 10k$)	28.4	20.8	62.6	11.77

Eval'06 (official results)	mWER[%]	mPER[%]	BLEU[%]	NIST
Baseline	40.6	28.7	50.0	10.80
+IBM+LM+Del+Length	40.4	28.5	50.9	10.92

Rescoring – Conclusion

- Some improvements for Spanish-English (Verbatim, FTE, ASR)
- Only modest results for English-Spanish:

  Verbatim: 45.2 $\rightarrow$ 45.4 BLEU%
  FTE: 49.1 $\rightarrow$ 49.4 BLEU%

  - Might be due to more complex morphology of the target language
- Experience shows that overfitting occurs when using too many features (i.e. no generalization on the test set)
- Most reliable: IBM model 1
- Good combination: IBM model 1 and additional LMs (preferably with larger n-grams than used for generation)
- Possible problem: lack of diversity in the N-best list (in contrast to system combination)
- Higher values for N only slightly decrease oracle ER, but introduce much more “noisy” hypotheses
- Manual comparison: hypotheses frequently differ in synonyms only

System combination

- Consensus translation can be computed by combining outputs of multiple systems
- Idea: select words which are present in the majority of translations (“voting”)
- Generate a possibly new translation
- To perform the voting correctly, a high-quality alignment of different hypotheses has to be determined
- Consider possible reordering of words/phrases

Idea of the algorithm

- Align different MT system outputs for each source sentence:
  - allow word reordering
  - take the context of the whole (test) document of translations into account
  - get a more reliable alignment by using an iterative alignment procedure
- Construct a confusion network from the (possibly reordered) translation hypotheses based on the alignment
- Use global system probabilities and other statistical models to select the best consensus translation from the confusion network

Alignment

Given a single source language sentence, combine M translation hypotheses from M translation systems:

- **choose one of the hypotheses E_m as the "primary" hypothesis, assume it to have correct word order**
- **align all other hypotheses $E_n (n = 1, ..., M; n \neq m)$ with E_m and reorder the words to match the word order of E_m**
- **repeat the procedure M times by letting each hypothesis play the role of the primary hypothesis once**

Alignment (contd)

- Alignment is performed in analogy to the training procedure in SMT (however, the sentences that have to be aligned are in the same language)
- Iterative unsupervised alignment training using the GIZA++ toolkit
- Pairwise alignment of the output of M systems for N test sentences ($N = 500 \ldots 2000$)
- Total size of the alignment training corpus is $M \cdot (M-1) \cdot N$ sentence pairs
- 4 iterations of IBM Model 1 and 5 iterations of the HMM model
- IBM Model 1 single-word lexicon probabilities are initialized
  - with co-occurrence counts of identical words in E_n and E_m
  - with fractions of a count for words with identical prefixes

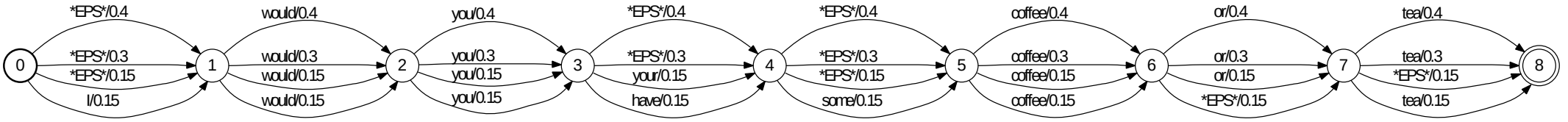
Reordering

- Reorder the sentence E_n based on the alignment with the primary hypothesis E_m
- Use the final HMM alignment that is a function of words in E_n
- The words of E_n are reordered based on this alignment, such that the final alignment between E_m and E_n becomes monotone
- Overall, determine $M-1$ monotone one-to-one alignments between E_m and E_n for $n = 1, ..., M; n \neq m$
- Construct a confusion network from these alignments

Building a confusion network

Example:

Original hypotheses	1. would you like coffee or tea 2. would you have tea or coffee 3. would you like your coffee or 4. I have some coffee tea would you like
Alignment and reordering	would|would you|you have|like coffee|coffee or|or tea|tea would|would you|you like|like your|$ coffee|coffee or|or $\|tea I\|$ would|would you|you like|like have|$ some|$ coffee|coffee $|or tea|tea
Confusion network	$ would you like $ $ coffee or tea $ would you have $ $ coffee or tea $ would you like your $ coffee or $ I would you like have some coffee $ tea



Extracting Consensus Translations

- Introduce global system probabilities
  - tuned manually based on the performance of the individual systems on a development set
- Perform "voting" on each of the M confusion networks:

0.25	$	would	you	like	$	$	coffee	or	tea
0.35	$	would	you	have	$	$	coffee	or	tea
0.1	$	would	you	like	your	$	coffee	or	$
0.3	I	would	you	like	have	some	coffee	$	tea
Voting	$/0.7 I/0.3	would/1.0	you/1.0	have/0.35 like/0.65	$/0.6 your/0.1 have/0.3	$/0.7 some/0.3	coffee/1.0	or/0.7 $/0.3	tea/0.9 $/0.1

- Unite M confusion networks into one automaton
- Extract consensus translation using
  - the single-best path or
  - N best paths for further processing (e.g. rescoring)

Translations of European Parliamentary Speeches

TC-STAR 2005 Evaluation, Spanish-English verbatim condition (case-insensitive evaluation, no punctuation):

EPPS Spanish-English	WER [%]	PER [%]	BLEU [%]
worst single system	49.1	38.2	39.6
best single system	41.0	30.2	47.7
consensus of 4 systems	39.1	29.1	49.3
+ rescoring	38.8	29.0	50.7

TC-STAR 2006 Evaluation, English-Spanish verbatim condition (case-sensitive evaluation with punctuation):

EPPS English-Spanish	WER [%]	PER [%]	BLEU [%]
worst single system	47.6	36.1	40.1
best single system	43.1	32.1	45.4
consensus of 5 systems	40.9	30.4	47.5

System combination – Conclusion

- Novel algorithm for computing consensus translations from the output of multiple MT systems
- The approach aligns the alternative translation hypotheses, allowing for word reordering
- The decision on how to align two translations of a sentence takes the whole document of translations into account
- Large and significant gains in translation quality obtained on different tasks and conditions
- Best translations in the TC-STAR 2006 MT evaluation according to all objective error measures
- The method can be applied when translating automatically transcribed speech to reduce the negative impact of speech recognition errors on translation accuracy

Conclusions

- Two approaches using multiple hypotheses for improving MT:
  - Rescoring: use N-best translations and apply reranking
  - System combination: compute consensus translations from different MT systems
- Some improvements for rescoring on EPPS task
- Good improvements for system combination:
  $\longrightarrow$ diversity of the various translations seems to be important
- Advantages of rescoring:
  - test new models easily (direct integration in the search process might be complicated and time-consuming)
  - apply models on the whole sentence level (structural properties, long-distance dependencies, grammar-based approaches)
- Methods can be combined: reranking an N-best list generated from a combination of systems yields additional improvements

Thank you for your attention

Saša Hasan

hasan@informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

References

[Bangalore & Bordel+ 01] S. Bangalore, G. Bordel, G. Riccardi: Computing Consensus Translation from Multiple Machine Translation Systems. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, 2001. 4

[Fiscus 97] J.G. Fiscus: A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–352, Santa Barbara, CA, 1997. 4

[Hasan & Bender+ 06] S. Hasan, O. Bender, H. Ney: Reranking Translation Hypotheses Using Structural Properties. In *EACL06 Workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy, 2006. To appear. 4

[Hasan & Ney 05] S. Hasan, H. Ney: Clustered Language Models based on Regular Expressions for SMT. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May 2005. 4

[Jayaraman & Lavie 05] S. Jayaraman, A. Lavie: Multi-Engine Machine Translation Guided by Explicit Word Matching. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, pp. 143–152, Budapest, Hungary, May 2005. 4

[Nomoto 04] T. Nomoto: Multi-Engine Machine Translation with Voted Language Model. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July 2004. 4

[Och 03] F.J. Och: Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, Sapporo, Japan, July 2003. 4

[Och & Gildea+ 04] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, D. Radev: A Smorgasbord of Features for Statistical Machine Translation. In *Proc. 2004 Meeting of the North American chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 161–168, Boston, MA, 2004. 4

[Och & Ney 02] F.J. Och, H. Ney: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 295–302, Philadelphia, PA, July 2002. 4

[Paul & Doi+ 05] M. Paul, T. Doi, Y. Hwang, K. Imamura, H. Okuma, E. Sumita: Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation. In *International Workshop on Spoken Language Translation*, pp. 55–62, Pittsburgh, PA, 2005. 4

[Shen & Sarkar+ 04] L. Shen, A. Sarkar, F.J. Och: Dicriminative Reranking for Machine Translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, Boston, MA, May 2004. 4

Corpus statistics

		Spanish	English
Train	Sentences	1 167 627	
	Words + Punct. Marks	35 320 646	33 945 468
	Words	32 074 034	30 821 291
	Vocabulary	159 080	110 636
	Singletons	63 045	46 121
Dev	Sentences	1 452	1 122
	Words + Punct. Marks	52 087	28 348
	Words	46 816	25 885
	Distinct Words	7 013	4 162
	OOV Words	351	93
Test	Sentences	1 782	1 117
	Words + Punct. Marks	56 468	28 492
	Words	50 634	25 869
	Distinct Words	7 204	4 172
	OOV Words	363	72

Translation examples – Effect of rescoring

Baseline	*has been distributed the final draft of the agenda of the plenary in June . . .*
Rescoring	*It has been distributed to the final draft of the agenda of the plenary in June . . .*
Reference	*The final project for the agenda of the plenary session of June was distributed . . .*
Baseline	*. . ., we are receiving very worrying news .*
Rescoring	*. . ., we are receiving very disturbing reports .*
Reference	*. . ., we are receiving very distressing news .*
Baseline	*We are facing a crisis whose emergence can not be seen, that some have referred of genocide, and which has caused, in any case, thousands of people dead . . .*
Rescoring	*We are facing a crisis whose emergence can not be seen, some have referred to as genocide, and which has caused, in any case, thousands of deaths . . .*
Reference	*We are facing a crisis, the exit of which is hard to see, which some branded as genocide, and which, in any case, caused thousands of dead . . .*
Baseline	*This proposal, for the first time, the co-financing of projects in the field of energy and not only the prior studies.*
Rescoring	*This proposal envisages , for the first time, the co-financing of projects in the field of energy and not only the prior studies.*
Reference	*Said proposal contemplates, for the first time, the co-financing of projects in the energy sector, and not only the preliminary surveys.*

Synonyms encountered (baseline / rescoring): *in this area / in this field*, *trust in / rely on*, *intolerable / inadmissible*, *ability / skill*, *appeared / emerged*, *jointly with / together in*, . . .

Translation examples – System combination

Best system	***I also authorised to committees to certain reports***
Consensus	***I also authorised to certain committees to draw up reports***
Reference	***I have also authorised certain committees to prepare reports***
Best system	***human rights which therefore has fought the european union***
Consensus	***human rights which the european union has fought***
Reference	***human rights for which the european union has fought so hard***
Best system	***we of the following the agenda***
Consensus	***moving on to the next point on the agenda***
Reference	***we go on to the next point of the agenda***