Introduction to Statistical Machine Translation *Fall 2007*

Instructor: Rafael E. Banchs (rbanchs@gps.tsc.upc.edu)

Office hours: Mondays, from 09:00 to 10:00, at D4-100

Meeting days: Mondays, from 10:00 to 11:00. Theory sessions.
(Mondays/Thursday): from 11:00 to 13:00. Lab sessions.

Course description:
This course introduces students to basic concepts in Statistical Machine Translation (SMT) from both theoretical and practical points of view.

1. The theory sessions include a brief review of machine translation history and a complete description of the state of the art. Different approaches to machine translation will be briefly revised, but particular attention will be paid to the statistical approach. Some of the key concepts in SMT, such as word alignment, model estimation, and decoding will be discussed in detail. These concepts and their related problems will be illustrated by using some simple exercises.
2. The lab sessions include some guided SMT exercises by using some available free-distribution software tools. Also, a short research project on SMT will be developed.

Objectives:

1. To learn about the state of the art in Statistical Machine Translation (SMT).
2. To become familiar with some on-line available resources for SMT research.
3. To develop a short research project in SMT.

Program:
Monday, NOV-05: Course introduction; machine translation, a general overview.
Monday, NOV-05: 1st lab session: a data-driven machine translation experience revisited.

Monday, NOV-12: Statistical machine translation basics and evaluation metrics.
Monday, NOV-12: 2nd lab session: statistical machine translation of european parliament data.

Monday, NOV-19: From word-based to phrase-based statistical machine translation.
Monday, NOV-19: 3rd lab session: alignment set and language model impact on translation quality.

Monday, NOV-26: Decoding and n-best list rescoring.
Monday, NOV-26: 4th lab session: log-linear weight optimization and phrase-based SMT features.

Monday, DIC-03: Application examples and the future of statistical machine translation.
Monday, DIC-03: 5th lab session: short research project.

Monday, DIC-10: Discussion session "Machine Translation and the Information Society".
Monday, DIC-10: 6th lab session: short research project.

Monday, DIC-17: Presentation of research projects.
Monday, DIC-17: 7th lab session: short research project.

Evaluation:

1. Individual exercises and homework assignments (30%)
2. Attendance (at least 80% of the sessions) and participation (20%)
3. Short research project and final report (50%)

Website: Additional and updated information is available at: http://gps-tsc.upc.es/veu/personal/rbanchs/smt/, you will need the following username "smt" and password "fall07" to be able to enter the site.